

An investigation into the clinical and genomic diversity of  
Human Papillomavirus in invasive disease in Scotland  
and the implications of new technologies

Daniel Guerendiain Regalado

A Thesis Submitted for the Degree of PhD  
at the  
University of St Andrews



2023

Full metadata for this thesis is available in  
St Andrews Research Repository  
at:

<http://research-repository.st-andrews.ac.uk/>

Identifier to use to cite or link to this thesis:

DOI: <https://doi.org/10.17630/sta/627>

This item is protected by original copyright

This item is licensed under a  
Creative Commons License

<https://creativecommons.org/licenses/by-nc-nd/4.0>

## **Candidate's declaration**

I, Daniel Guerendiain Regalado, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 48,957 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in October 2018.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 3<sup>rd</sup> May 2023

Signature of candidate

## **Supervisor's declaration**

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 3<sup>rd</sup> May 2023

Signature of supervisor

## **Permission for publication**

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Daniel Guerendiain Regalado, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

**Printed copy**

No embargo on print copy.

**Electronic copy**

No embargo on electronic copy.

Date 3<sup>rd</sup> May 2023

Signature of candidate

Date 3<sup>rd</sup> May 2023

Signature of supervisor

## **Underpinning Research Data or Digital Outputs**

### **Candidate's declaration**

I, Daniel Guerendiain Regalado, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date 3<sup>rd</sup> May 2023

Signature of candidate

### **Permission for publication of underpinning research data or digital outputs**

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date 3<sup>rd</sup> May 2023

Signature of candidate

Date 3<sup>rd</sup> May 2023

Signature of supervisor



## General Acknowledgments

I would like to thank my supervisor Professor Matthew Holden for his support, tutelage, invaluable patient and feedback during the course of my PhD degree. A special thanks to Dr Kate Cuschieri, without whom I would not have made it through my PhD degree. I will never forget your tutelage, advice, and your patient over these years. Thanks for always find time for me.

I would like to offer my special thanks to Dr Ingólfur Johannessen and Professor David Harrison for making this PhD happen. I will always remember the opportunity you presented me in 2016. Additionally, this endeavour would not have been possible without the generous support from NHS Lothian, who financed my research.

My gratitude extends to the Scottish HPV Reference laboratory for their advice, helping finding samples and the long lab conversations. And for allowing me to use your reagent and equipment! And to SERL, for being extremely patient with me and my lack of sleep during the last months of the writing.

Thank you to my parents for all their effort and sacrifices they made for me. Thanks to them I have been able to complete this journey and become a scientist.

And my biggest thanks to my wife Marta and daughter Carla, for all the support and strength you have shown me through this stage. Your belief has kept my spirits and motivation high during all these years. . I could not have done it without your help.

## **Funding**

This work was supported by the Department of Laboratory Medicine, NHS Lothian

## **Research Data/Digital Outputs access statement**

Research outputs underpinning this thesis are available at:

<https://doi.org/10.3390/diagnostics12123222>

<https://doi.org/10.1002/cam4.4771>

<https://doi.org/10.3390/v13071323>

## Abstract

This study investigated the prevalence of human papillomavirus (HPV) in different HPV-driven cancers in Scotland, including cervical, oropharyngeal, and anal. The study found that 91.5% of cervical, 55% of oropharyngeal and 88.6% of anal cancers are driven by HPV. Data has shown that most of these infections could potentially be prevented by the HPV vaccines.

This research also investigated novel molecular methods for detecting HPV, such as droplet digital PCR (ddPCR) and next-generation sequencing (NGS), which can provide more comprehensive data about the virus and its various (sub)-lineages. ddPCR was used to determine the number of HPV 16 copies per cell in anal cancers. It was found that the qualitative presence of HPV and high viral load was associated with more prolonged survival in anal cancers, consistent with other HPV-driven cancers.

Prior to the sequencing, three different extraction methods were examined to identify which one could be best for NGS downstream. Additionally, a bioinformatics pipeline was set up and validated by comparing results with the International HPV Reference Centre.

NGS was used to analyse the genome of the HPV 16 detected in anal cancer samples and anal swabs from an asymptomatic male population. By analysing the whole genome of the virus, HPV 16 sub-lineages were identified. Lineages A1 and A2 were the most prevalent in both groups, and only minimal differences were detected for sub-lineages B1, C1, and D1. Analysing sub-lineages and clinical data showed no overall survival differences between A1 and non-A1 sub-lineages.

In addition, the potential of NGS for HPV diagnosis, compared with conventional HPV testing and how NGS may be applied for the detection and risk stratification for HPV infection and associated diseases was also described. Finally, a case scenario was presented for guidance in implementing direct sequencing in an HPV reference laboratory.

## Lay Summary

Human papillomavirus (HPV) is the most frequent sexually transmitted infection; most humans will be infected at some point in their life. However, not all infections are the same. Persistent HPV infections could lead to pre-cancerous lesions and, if not cleared, cancer. There are more than 200 HPV types that can infect humans. HPV type 16 (HPV 16) is the most associated HPV type with cancer. However, 13 other high-risk types are also causative of cancer in different anatomical locations, including the cervix, head and neck, anus, vulva, vagina, and penis. HPV vaccine programmes have been introduced in several countries since 2007, effectively preventing infection and associated diseases.

Although HPV infections tend to clear after a couple of years, a persistent infection is the leading cause of the lesion's malignancy. Thus, diverse screening programs have been established in different countries, aiming to detect HPV before developing pre-cancer and cancer lesions.

Cervical cancer is an HPV-driven lesion with higher incidence worldwide, and most of the screening programs available worldwide are focused on cervical cancer prevention. For the rest of HPV-driven cancers, the lower incidence in the population, very few screening programs are in place, leading to a shortage of knowledge about the HPV types present in oropharyngeal and anal cancers, including the vaccine-preventable fraction.

In Scotland, anal cancer incidence has increased over the years. A cancer whose treatment has a significant impact on a patient's life. However, HPV testing is not routinely performed in anal lesions, and there is no screening program in place to detect HPV presence in anal lesions (pre-cancerous or cancer lesions)

This PhD thesis has analysed the positivity of HPV in cervical (2015 – 2017), oropharyngeal (2013 to 2020), and anal cancers (2009 to 2018) samples collected in Scotland. It was identified that HPV drives 91.5% of cervical cancers and that the HPV vaccine could potentially prevent 87.7%. For the oropharyngeal cancer group, almost 55% were potentially caused by HPV, with almost all vaccine-preventable. In anal cancers, 88.6% of samples were positive for at least 1 HPV type, 98% of the HPV-positive cases being potentially vaccine-preventable. Additionally, thanks to the clinical information obtained from the anal cancer patients, it was identified that those HPV-positive anal cancers had improved overall survival compared with HPV-negative anal cancers.

The technology used to diagnose microorganisms in infections, including HPV, has dramatically evolved in the last few years. PCR has been established as the gold standard test, but new molecular techniques like droplet digital PCR (ddPCR) and next-generation sequencing (NGS) have recently been explored. The increment in the use of NGS has been mainly associated with the reduction in cost and easier access to sequencing instruments. Both technologies supplement the data/information obtained from a conventional HPV test. The high sensitivity of droplet digital PCR allows the detection and quantification of the number of copies of the HPV per cell, while next-generation sequencing allows one to "read" the entire genome and identify minor differences within types, SNPs, or missing regions in the viral genome.

Thanks to ddPCR, it was possible to determine the number of HPV 16 copies per cell (viral load) present in the HPV 16-positive anal cancers. Access to a group of anal cancer on which clinical and survival information was known allowed me to investigate the association between viral load and overall survival. A high viral load has been associated with more prolonged survival in oropharyngeal cancer compared with a low viral load. Furthermore, by analysing the association of viral load with survival in anal cancer

patients, it was identified that this also applies to anal cancer. Those anal cancers with high viral load had a more prolonged survival than those with a low number of virus copies.

Even if the HPV family is classified in types, within these types, there exist differences in the genome that allow categorising them in lineages (differences between 1 – 10%) and sub-lineages (0.5 to 1%). Despite this small difference, some of these lineages or sub-lineages have been associated with higher persistent infections or higher risk of developing cancer, particularly cervical cancer. In the present work, NGS was used to analyse the genome of the HPV 16 detected in anal cancer samples and compared it with the HPV 16 genome detected in anal swabs from an asymptomatic population.

However, before starting with the NGS, due to the novelty of NGS in HPV diagnosis, it was necessary to distinguish what was the best nucleic acid extraction methods to obtain the highest quality and quantity of DNA from the cancer specimens. Additionally, it was necessary to develop a protocol that fitted with the capacity of the laboratory, sample type and the information we wanted to obtain. Also, it was necessary to prepare and validate a protocol to analyse the data obtained from the sequencing.

Once the NGS protocol was defined, whole genome sequencing was performed using the MiSeq platform. Analysing the whole genome of the HPV 16 in anal cancers identified the sub-lineages present in both anal cancer and the asymptomatic cohort. Most of the sub-lineages detected belong to lineage A (A1 and A2 mainly), not detecting enough other sub-lineages to evaluate the sub-lineage influence on overall survival. In the final part of this thesis, the potential applications for NGS for HPV diagnostics, and how they compare with conventional HPV tests were described. In addition, a case study has been presented describing step by step how to implement a direct sequencing approach into an HPV reference laboratory.

In conclusion, this study has found that an 87% of cervical, 52% of oropharyngeal and 88%% of anal cancers can be potentially prevented by the HPV vaccine. New technologies like ddPCR and NGS enhance PCR diagnosis and offer an opportunity to not only look at the presence or absence of HPV but also to determine viral load levels or identify minor differences in the viral genome that could impact the survival of the patient or response to treatment.

## List of figures

Figure 1. Estimated number of new and existing STI in the USA .....	22
Figure 2. Structure and organisation of HPV 16 genome.....	24
Figure 3. Cervical cancer incidence in Scotland from 1982 - 2016.....	41
Figure 4. Oropharyngeal cancer incidence (European Age-specific rates) Scotland: 1993 – 2017 .....	42
Figure 5. Anal cancer incidence (European Age-standardises rates), Scotland 1993 - 2017 .....	43
Figure 6. Example of Anyplex II HPV28 report .....	58
Figure 7. Diagram describing both Scottish cervical biopsy surveillance studies.....	65
Figure 8. EASR of the 5 most common HPV-driven cancers in Scotland.....	81
Figure 9. Incidence of HPV-driven cancers detected in Scotland in 2017.....	83
Figure 10. Number of cervical cancer samples collected in each Scottish NHS healthboards.....	87
Figure 11. NHS Health board distribution of oropharyngeal samples received between 2009 – 2018. ....	89
Figure 12. HPV type prevalence by HPV type and morphology .....	93
Figure 13. HPV type prevalence in 2015, 2016 and 2017 for all cervical cancer samples.....	96
Figure 14. HPV type prevalence in 2015, 2016 and 2017 for SCC samples.....	96
Figure 15. HPV type positivity in 2015, 2016 and 2017 for ASC + ADC samples.....	97
Figure 16. Changes in HPV positive, HPV-negative, HPV 16-positive and HPV 16 and/or HPV 18-positive prevalence in the different age groups.....	98
Figure 17. HPV type prevalence by HPV type – Oropharyngeal I cancer samples .....	100
Figure 18. HPV type prevalence in oropharyngeal cancer samples from 2013 to 2018 .....	102
Figure 19. HPV positivity by age group for oropharyngeal cancer samples.....	103
Figure 20. Age Standardized per 100,000 incidence and mortality rates for anal cancer worldwide and by regions.....	111
Figure 21. European Age-Standardised Incidence Rates per 100,000 population, Scotland .....	112
Figure 22. Number of new cases of anal cancer by age-specific group in per year in the UK.....	115
Figure 23. Overview of the process followed including HPV genotyping of 221 anal samples.....	117
Figure 24. HPV Prevalence in 200 anal lesions (AIN + cancer). HPV Prevalence in 200 anal lesions .....	122
Figure 25. HPV positivity in 185 anal cancer samples collected between 2009 to 2018 .....	123
Figure 26. Overall Survival probability for “any” HPV positive and HPV negative and for HPV 16-positive vs. HPV-negative anal cancer cases using Kaplan Meier estimator.....	128
Figure 27. Kaplan-Meier survival curve stratified by viral load (Low and Medium/High).....	130
Figure 28. Boxplot - Quantity of DNA obtained from the three different extraction methods .....	142
Figure 29. Boxplot - Ct value obtained from the three extraction methods.....	143



Figures 30 - 33. Tapestation graph with the sample intensity (FUI) for the DNA library's different sizes 145-142

Figure 34. Phylogenetic tree representing the HPV 16 sub-lineages present in the anal sample cohort based on core SNPs ..... 162

Figure 35. Prevalence of HPV 16 sub-lineages in anal cancer and control cohort. .... 164

Figure 36. Kaplan-Meier survival curve stratified by HPV 16 sub-lineages (A1 vs non-A1). .... 167

Figure 37. Qualimap Plot showing an anal cancer sample with no integration ..... 169

Figures 38 - 50. Examples of potential Integration of HPV genome ..... 170- 167

Figure 51. Diagram of the HPV NGS workflow for whole genome sequencing or target sequencing. .... 185

Figure 52. Description of the necessary steps to perform a direct sequencing approach for HPV investigation. .... 201

## List of tables

Table 1. Classification of the main HPV types according to their association with cancer, lineage, and sub-lineage.	27
Table 2. Number of all cancer cases attributable to HPV and corresponding attributable fraction (AF) for all cancers, cancer site(s) and sex.....	40
Table 3. Number of different cancer cases attributable to HPV for various locations .....	40
Table 4. Trend in coverage of first dose of HPV immunisation by the end of school years 2014/15 - 2021/22 in Scotland .....	45
Table 5. Consumables and Manufacturer.....	52
Table 6. Equipment and Manufacturer.....	53
Table 7. Seegene Anyplex 28 II Melting PCR cycles. ....	57
Table 8. Characteristics of the 3 nucleic acid methods examined in chapter 5.....	<b>Error! Bookmark not defined.</b>
Table 9. Name and sequence of HPV 16 primers and probe. ....	62
Table 10. Total number of oropharyngeal samples collected between 2013 and July 2020.....	66
Table 11. Primers pools for HPV 16 sequencing. ....	70
Table 12. Total number of high-grade and cervical invasive lesions collected, stratified by morphology. ....	85
Table 13. Total number of oropharyngeal samples collected between 2013 and July 2020 .....	88
Table 14. High risk HPV types in SCC and ASC + ADC in the cervix .....	94
Table 15. Most prevalent HPV types on all cervical cancer samples (irrespective of morphology) by year.....	94
Table 16. HPV types detected in oropharyngeal cancer samples between 2013 and 2020. ....	101
Table 17. Demographic & clinical characteristics of the anal samples collected between 2009 to 2018 in the South-East of Scotland.. ....	118
Table 18. Demographic & clinical characteristics of the anal cancer samples collected between 2009 to 2018 in the South-East of Scotland.....	125
Table 19. Univariate and multivariate hazard ratio of HPV status derived using Cox regression for anal cancer cases. ....	128
Table 20. Level of viral loads by vital status obtained in the HPV 16-positive group anal cancer cases. ....	129
Table 21. Overall survival stratified by clinical variables, demographic variables and L1 viral load. HR derived using Cox regression. ....	131
Table 22. Minimum, maximum and average values were obtained from Qubit extraction and qPCR for the three different extraction methods. ....	141
Table 23. Comparison of HPV 16 sub-lineage identified by Karolinska Institute.....	148
Table 24. HPV 16 sub-lineages breakdown in the anal cancer cohort.....	161
Table 25. HPV 16 sub-lineages breakdown in the asymptomatic population (rectal swabs) .....	163
Table 26. HPV 16 sub-lineage distribution in males and females in the anal cancer cohort. ....	165

Table 27. A1 positivity status according to demographics and clinical variables. ....	166
Table 28. A univariate and multivariate hazard ratio of HPV 16 sub-lineages derived using Cox regression.....	168
Table 29. Missing genes in the suspect samples of HPV being integrated. ....	169
Table 30. Demographics of anal cancers with a plausible integration. ....	175
Table 31. The number of results obtained in PubMed. Search performed in PubMed on the 2 <sup>nd</sup> of March 2023 ....	182
Table 32. The cost associated with an NGS genotyping approach (not WGS).....	186
Table 33. The cost associated with NGS by PCR target enrichment (WGS).....	187
Table 34. The cost associated with direct NGS (24 samples). Includes reactions for extraction, library prep and sequencing (MiSeq). Cost of reagents obtained in December 2022. ....	189
Table 35. The cost associated with long read sequencing (ONT). Includes reactions for extraction, library prep and sequencing (Nanopore) for 24 samples. Cost of reagents obtained in December 2022. ....	191
Table 36. Cost per sample for the NGS approaches described.....	191
Table 37. Advantages and Disadvantages of each of the NGS approaches.....	193
Table 38. Best HPV test by objective (conventional vs NGS). ....	197

## List of abbreviations

- AC anal cancer
- ADC adenocarcinomas
- AF attributable fraction
- AIN anal intraepithelial neoplasia
- AIS adenocarcinoma in situ
- AJCC American Joint Committee on Cancer
- ANOVA analysis of variance
- ASC adenosquamous carcinoma
- ASCC anal squamous cell carcinoma
- ASR age standardised rate
- CCX cervical cancer
- CI or Cis confidence interval
- CIN cervical intra-epithelial neoplasia
- CRT chemoradiotherapy
- ctDNA circulating DNA
- DNA deoxyribonucleic acid
- ds double stranded
- E6AP E6-associated protein
- EASR European age-standardised incidence rate
- ESMO European society for medical oncology
- FIGO International Federation of Gynaecology and Obstetrics
- FFPE formalin-fixed paraffin embedded
- GGC Great Glasgow and Clyde health board
- GUM genitourinary medicine
- HIC high-income countries
- HIV human immunodeficiency virus
- H&S head and neck
- Tukey HSD Tukey honestly significant difference
- HPV human Papillomavirus

- HPV+ve human Papillomavirus positive
- HPV-ve human Papillomavirus negative
- hr-HPV high risk human papillomavirus
- HR hazard ratio
- IARC International agency for research on cancer
- ICC invasive cervical cancer
- IFN interferons
- IQR interquartile range
- ISD Information Services Division (Scotland)
- IVD in vitro diagnostics
- JCVI Joint Committee on Vaccination and Immunisation
- KM Kaplan-Meier estimator
- LBC liquid-based cytology
- LMIC low and middle-income countries
- LR-HPV low risk human papillomavirus
- mRNA messenger RNA
- MSM men who have sex with men
- NA nucleic acid
- N/A not available
- NGS next generation sequencing
- NHS National Health Service
- NRS NHS Research Scotland
- ONT Oxford Nanopore technologies
- OPC oropharyngeal cancer
- OPSCC oropharyngeal squamous cell carcinoma
- OR odds ratio
- ORF open reading frame
- OS overall survival
- PAVe papillomavirus episteme
- PCR polymerase chain reaction
- PeIN penile intraepithelial neoplasia

- PHS            Public Health Scotland
- PV            papillomavirus
- RNA          ribonucleic acid
- SBSTRL      Scottish bacterial sexually transmitted infections Reference Laboratory
- SCC          squamous cervical carcinoma
- SD            standard deviation
- SERL        Scottish *E. coli* Reference Laboratory
- SHPVRL     Scottish HPV Reference Laboratory
- SMRL        Scottish Mycobacteria Reference Laboratory
- SNP          single nucleotide polymorphism
- SSVC        Scottish Specialist Virology centre
- TNM          tumour, node, metastasis
- UK            United Kingdom
- UKAS        United Kingdom accreditation Service
- URR          upstream regulatory region
- US or USA    United States of America
- VAIN        vaginal intraepithelial neoplasia
- VIN          vulval intraepithelial neoplasia
- VL            viral load
- WGS        whole genome sequencing
- WHO        World Health Organization

# Table of Contents

Acknowledgment .....	5
Abstract .....	7
List of figures .....	12
List of tables .....	14
List of abbreviations .....	16
Table of Content .....	19
1. Introduction .....	22
1.1 Papillomavirus .....	23
1.2 Structure of the HPV genome .....	23
1.3 The biology and life cycle of HPV .....	24
1.4 Host response to infection .....	26
1.5 HPV classification attending to carcinogenic risk .....	26
1.6 HPV lineages, sub-lineages and oncogenic risk associated .....	29
1.7 Persistent HPV lesions and pathology classification .....	33
1.8 HPV detection – conventional tests .....	35
1.9 Next generation sequencing and whole genome sequencing .....	36
1.10 Burden of diseases driven by HPV .....	38
1.11 Disease management in Scotland: vaccination and screening .....	43
1.12 Epidemiology of HPV in Scotland .....	45
1.13 HPV screening and surveillance in Scotland .....	48
1.14 Role of the Scottish Human papillomavirus Reference Laboratory .....	49
1.2 Aims of the thesis .....	50
2. General Materials and Methods .....	52
2.1 Materials .....	52
2.2 Equipment .....	53
2.3 Nucleic acid extraction and HPV genotyping .....	54
2.4 Governance .....	63
2.5 Information capture for cervical and oropharyngeal cancer data .....	64
2.6 Droplet digital PCR general steps .....	68

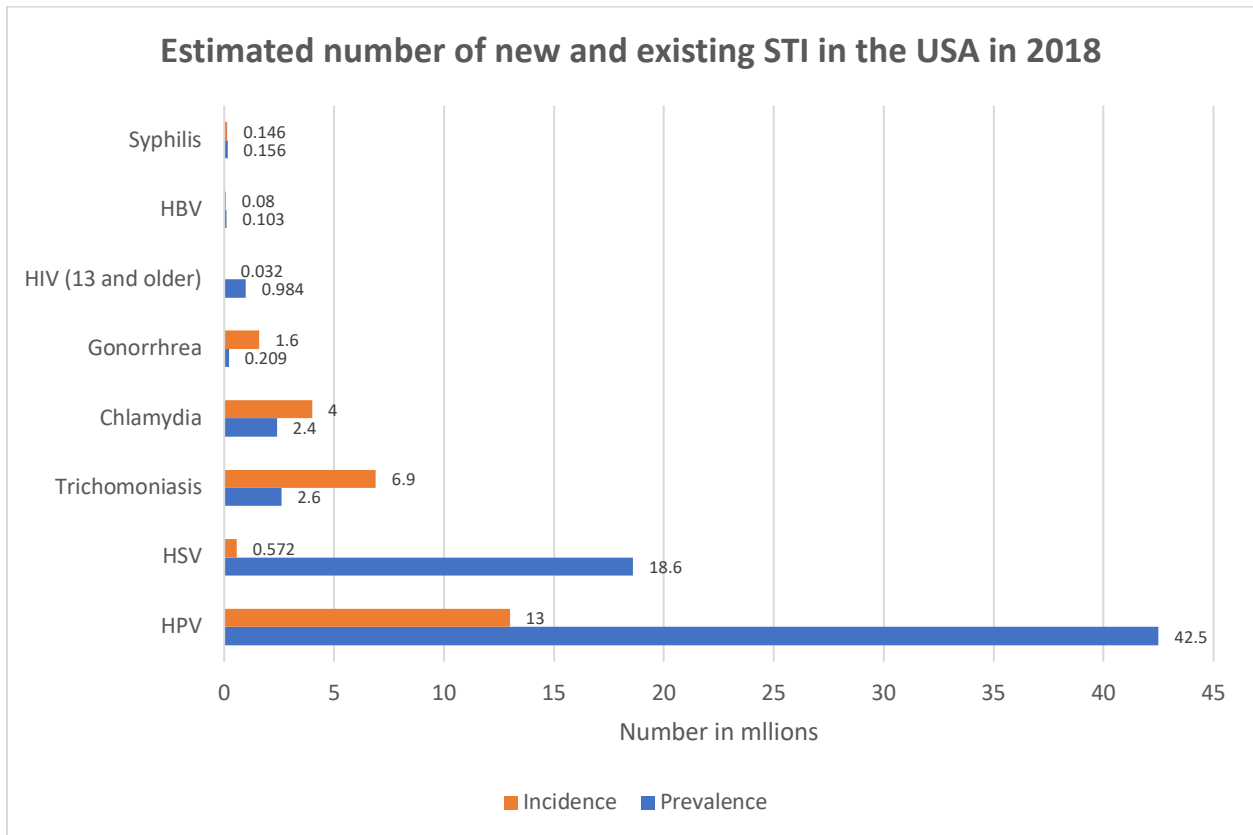
2.7 General NGS methods used in chapter 6.....	69
2.7.5 Bioinformatic Analysis.....	72
2.7.6 Validation of the library preparation and bioinformatic analysis.....	74
2.8 Statistical packages.....	74
2.9 Primers.....	75
<b>3. HPV type specific prevalence in cervical and oropharyngeal cancer in the Scottish population.....</b>	<b>81</b>
3.1 Introduction.....	81
3.2 Overarching Aim:.....	84
3.3 Material and Methods.....	85
3.4 Results.....	91
3.5. Discussion.....	103
<b>4. Role and influence of HPV in anal disease in the South-East of Scotland. HPV type specific prevalence and viral load.....</b>	<b>109</b>
4.1 Introduction.....	109
4.2 Material and Methods:.....	117
4.3 Results.....	121
4.4 Discussion.....	132
<b>5. Optimisation of molecular and bioinformatic tools to support the identification of HPV 16 sub-lineages.....</b>	<b>136</b>
5.1 Introduction.....	136
5.2 Material and Methods:.....	139
5.3 Results:.....	141
5.4 Discussion.....	148
<b>6. Identification of HPV 16 sub-lineages, association with demographics and clinical variables and overall survival in anal cancer and asymptomatic cohort.....</b>	<b>151</b>
6.1 Introduction.....	151
6.2 Material and Methods.....	156
6.3 Results.....	160
6.4 Discussion.....	175
<b>7. Applications and translation of next generation sequencing.....</b>	<b>179</b>
7.1 Introduction.....	179
7.2 Material and Methods.....	183
7.3 Results.....	184



7.4 Discussion .....	202
8. Final Discussion .....	<b>205</b>
9. References .....	<b>214</b>
10. Appendixes .....	2304
11. Publications .....	<b>232</b>
12. Ethical Approval Documents .....	<b>273</b>

# 1. Introduction

Human papillomavirus (HPV) is one of the most common sexually transmitted infections (STI) in the world. Depending on the virus type and the anatomical site, infection can cause different clinical manifestations, from skin and genital warts to cancer<sup>1-3</sup>.



**Figure 1. Estimated number of new and existing STI in the USA.** Information collected from [cdc.gov](https://www.cdc.gov)<sup>4</sup>

HPV isolates are described as types and classified taxonomically<sup>2,5</sup> and according to the association with cancer. Those types frequently associated with cancer are classified as high-risk types and those which are found mainly in genital warts are classified as low-risk HPV types<sup>6</sup>.

Persistent infection with high-risk human papillomavirus (hr-HPV) has been identified as causative of almost all cervical cancer cases (99%)<sup>7</sup> as well as other anogenital and (a component of) head and neck cancers. HPV type positivity varies among different anatomical locations, for example HPV type 16 (HPV 16) is causative for ~90 – 95% of HPV-driven oropharyngeal<sup>8</sup> and anal cancer, while in cervical cancer, HPV 16 has been detected over 60% of the cases.<sup>8</sup>

## 1.1 Papillomavirus

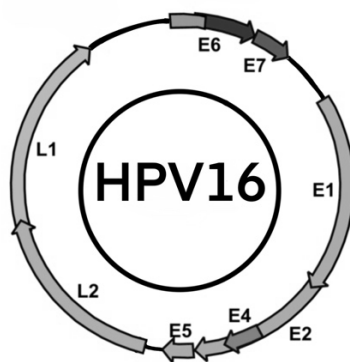
Papillomaviruses (PV) are small non-enveloped viruses with circular double-stranded deoxyribonucleic acid (DNA) genomes composed of approximately 8000 nucleotides. Papillomaviruses are highly species-specific viruses and have co-evolved with their host since their origin, for hundreds of millions of years. *Papillomaviridae* viruses has been detected in skin and mucosa of a variety of animal groups, including mammals, reptiles, and birds<sup>9</sup>. In humans, exposure to HPV is very common, and an estimated 65-100% of sexually active adults are exposed to HPV at different anatomic sites (oral, genital, or anal) at some point in their lifetime<sup>10</sup>. The infection is transmitted through mucosa and cutaneous skin by skin-to-skin contact. Age at first intercourse and number of sexual partners has been shown to influence acquisition rates of HPV infection both in women and men<sup>11</sup>.

## 1.2 Structure of the HPV genome

The HPV genome can be divided into 3 general regions: an upstream regulatory region (URR), an early region (E) and a late region (L) (Figure 2). The URR is noncoding and contains sequences that control viral transcription and replication. The early region, which contains open reading frames (ORFs) E1, E2, E4, E5, E6 and E7 is involved in multiple functions including the activation of transcription, transformation, replication, and viral adaptation to different conditions. The late region ORFs, encodes for the L1 and L2 capsid proteins which form the structure of the virion and facilitate viral DNA packaging and maturation

(Figure 2). However not all HPV types contain all these genes. Some HPV types (low risk types) do not have E6/E7-like functions and therefore they are not involved in the development of neoplastic lesions<sup>12-14</sup>. Over-expression of oncoproteins E6 and E7 is critical and necessary for HPV related progression to cancer<sup>15</sup>.

Even though the HPV genome (including for high-risk types) is composed of 8 genes, in some cancer samples some regions are not expressed or are lost. During the integration process of the viral genome into the host, loss of genetic segments occurs frequently. This is more likely to be found in carcinogenic lesions, as it is associated with higher risk of cancer<sup>16</sup>.



**Figure 2. Structure and organisation of HPV 16 genome.** From RD Burk *et al.* 2009.

### 1.3 The biology and life cycle of HPV

The life cycle of HPV begins with the virus entering the body often through a cut in the epithelial basal layer. The virus entry site seems to occur via binding to the heparan sulfate proteoglycans in the epithelial basement membrane via L1 and the virus is then taken up by endocytosis. Once inside the human cell, the viral genome enters to the cell nucleus through the nuclear pores or after the mitosis process<sup>17,18</sup>. Upon nuclear entry, E1 and E2 expression are associated with an initial phase of genome amplification, keeping the level of viral episome low (free viral DNA within the cell nucleus)<sup>19,20</sup>. In this stage, HPV is

capable of maintaining infection in the epithelial cells over time with a low expression<sup>21</sup>, which could be the explanation to the low immune response at this stage<sup>22</sup>.

HPV viral expression is carried out while infected basal cells differentiate and move up within the tissue layers<sup>21</sup>. E7 protein expression disrupts the interaction between Rb (Retinoblastoma protein) and E2F, resulting in the release of E2F factors in their transcriptionally active forms, stimulating the replication and cell-division<sup>23</sup>. As described in the Graham *et al.*, 2017, cells normally respond to any unexpected proliferation event by activating the apoptosis route. The HPV E6 protein has the ability to attach itself to two important proteins, E6-associated protein (E6AP) and p53, which plays a crucial role in apoptosis regulation, perturbing the control of cell cycle progression, leading to an increase of tumour cell growth, inhibition of the cell differentiation and induction of chromosomal instability, which may lead into tumourigenesis<sup>21</sup>.

In the late phase of the HPV cycle, E7, E1, E2, E4 and E5 proteins are expressed. E1 is a transcription factor, helicase activity and mediates in the episomal DNA replication. E2 is a transcription factor, regulating the viral copy number. E4 facilitates virion release. E5 stimulates cell proliferation and prevents differentiation<sup>21</sup>. When the epithelial cell is reaching the top layers, capsid proteins start to express L1 and L2 proteins, associated with the capsid proteins and assembly of the viral genomes. When fully formed, viral particles are released from the epithelial cell, infecting adjacent cells. In high grade lesions, viral gene expression is deregulated (denominated abortive infection). The gene expression tends to occur in disorder, detecting an abundant expression of E6, E7 through the epithelial layers and a low level of L1 and L2<sup>24</sup>.

#### **1.4 Host response to infection**

As described in the life cycle section, HPV cell infection can cause imbalance in the cell cycle, which could lead into carcinogenesis. To achieve this, HPV can evade the host immune system, remaining undetectable for long periods due to low number of copies of the virus in the infected cell<sup>25,26</sup>. Proteins E6 and E7 are the main proteins associated with the evasion of the immune response, starting by inhibiting the interferon synthesis<sup>27</sup>. They are also responsible for modifying the expression of cytokines, changing how antigens are presented, reducing the activity of IFN-pathways, and lowering the expression of adherence molecules<sup>27</sup>. This reduces the inflammatory response, reducing the presence of activated T-cells<sup>27</sup>. This ability of HPV to evade the immune response is crucial for the virus to successfully establish an infection.

#### **1.5 HPV classification attending to carcinogenic risk**

HPV infections can cause different clinical manifestations, from warts to neoplastic lesions in the different anatomical sites described before<sup>3,28,29</sup>. Up to 90% of persons infected with HPV clear the infection within about two years. However, the small part of the population that do not clear the infection are at a higher risk of progression to malignancy<sup>30</sup>.

The International Agency for Research on Cancer (IARC) evaluated the carcinogenic risk to humans of HPV infection<sup>1</sup>. The analysis considered results from more than 100 epidemiological studies, comprising case-control and cohort studies. These studies encompassed research conducted in the general population, as well as investigations targeting specific populations, such as transplant recipients and individuals with HIV, to explore the association between HPV and cancer. The review of these studies strongly supports the conclusion that certain types of HPV are responsible for causing cancer in humans. HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 66 are considered as carcinogenic to humans and belong to Group 1.

HPV 6 and HPV 11 are not carcinogenic to humans and belong to Group 2B. Classification and carcinogenic risk assigned can be found in Table 1.

**Table 1. Classification of the main HPV types according to their association with cancer, lineage, and sub-lineage.** Table prepared from data from De Villers *et al.* Vaccine, 2010<sup>13</sup> and Burk *et al.* Vaccine, 2013<sup>41</sup>.

Type	Carcinogenic	Lineage	Sub-lineage
6	Unlikely	A	
		B	B1, B2, B3
11	Unlikely	A	A1, A2
16	Yes	A	A1, A2, A3, A4
		B	B1, B2, B3, B4
		C	C1, C2, C3, C4
		D	D1, D2, D3, D4
18	Yes	A	A1, A2, A3, A4, A5
		B	B1, B2, B3
		C	
26	Probably	A	
31	Yes	A	A1, A2
		B	B1, B2
		C	C1, C2, C3
33	Yes	A	A1, A2, A3
		B	
		C	
35	Yes	A	A1, A2
39	Yes	A	A1, A2
		B	
40	Unlikely		
42	Unlikely		
43	Unlikely		
44	Unlikely		
45	Yes	A	A1, A2, A3
		B	B1, B2

51	Yes	A	A1, A2, A3, A4
		B	B1, B2
52	Yes	A	A1, A2
		B	B1, B2
		C	C1, C2
53	Probably	A	
		B	
		C	
		D	D1, D2, D3, D4
54	Unlikely	A	
		B	
		C	
55	Unlikely		
56	Yes	A	A1, A2
		B	
58	Yes	A	A1, A2, A3
		B	B1, B2
		C	
		D	D1, D2
59	Yes	A	A1, A2, A3
		B	
61	Unlikely	A	A1, A2
		B	
		C	
62	Unlikely		
64	Probably		
66	Probably	A	
		B	B1, B2
67	Probably	A	A1, A2
		B	
68	Probably	A	A1, A2
		B	
		C	C1, C2
		D	D1, D2
		E	
		F	F1, F2
69	Probably	A	A1, A2, A3, A4
70	Possibly	A	



		B	
71	Unlikely		
72	Unlikely		
73	Probably	A	A1, A2
		B	
74	Unlikely		
81	Unlikely	3	
82	Probably	A	A1, A2, A3
		B	B1, B2
		C	C1, C2, C3, C4, C5
83	Unlikely		
84	Unlikely		

## 1.6 HPV lineages, sub-lineages and oncogenic risk associated

### 1.6.1 HPV lineages and sub-lineages identification

As described before, Papillomavirus variants are genetically closely related, with above 99% nucleotide identity<sup>15</sup>. Nowadays, new data is emerging from epidemiological studies on the influence of various HPV variants (mostly HPV 16 and 18) on persistence and oncogenic risk thanks to the use of next generation sequencing technologies (NGS).

In the past, identification of lineages and sub-lineages was done by geographical location (European, African, Asian-American)<sup>5</sup>. In 2013, Burk *et al.* published an extensive study presenting all the lineages and sub-lineages of the main HPV types found in human samples and proposed a new way of identification/classification based on nucleotide identity differences of 1.0–10.0% and 0.5–1.0% between types, respectively<sup>15</sup>. The majority of the HPV types and lineages and sub-lineages associated with carcinogenic lesions are described in **Table 1**. HPV type 16 can be divided into four main variant lineages (A/B/C/D) and nine sub-lineages<sup>13</sup>: A, including A1-A3 (previously named European) and A4 (Asian) sub-lineages; B, including B1 (African-1, Afr1a) and B2 (African-1, Afr1b) sub lineages; C (African-2, Afr2a); and

D, including D1 (North American, NA1), D2 (Asian-American, AA2), D3 (Asian-American, AA1) and D4 sub-lineages. HPV type 18 can be divided into three major lineages (A, B, and C) and additional sub lineages (A1 to A5 and B1 to B3) that can be translated to the previous nomenclature (A1 and A2, = Asian American; A3 to A5, = European; and B/C, = African).

Recently, investigators have looked at sub-lineages present in distinct groups and their association with pre-cancer/cancer lesions. Mirabello *et al.* (2015) evaluated the association between HPV 16 lineages and risk of precancer/cancer in 3200 women from a US cohort<sup>31</sup>, using a whole genome sequencing (WGS) assay optimized for HPV genome sequencing<sup>32</sup>. This study confirmed the early observation of a higher risk of cervical precancer/cancer associated with B/C/D as a group compared to A lineages. They performed a case-control analysis (controls being HPV 16-positive women <CIN2 and cases being CIN2, CIN3, SCC, AIS, and adenocarcinoma). They found an overall association between HPV 16 lineage/sub-lineage and cervical pre-cancer/cancer risk. In addition, this study confirmed the earlier observation that some variants present a higher carcinogenic effect in women whose genetic background corresponds to that of the virus. In 2019, Clifford *et al* presented a study where they sequenced the whole genome of 7116 HPV 16 positives cervical samples from international epidemiological studies<sup>33</sup>. They used NGS for the identification of the variants and sub-lineages present in different global locations. They found that A1 sub-lineage was the most globally widespread and that sub-lineages had a strong regional specificity (A3 and A4 for East Asia, B1–4 and C1–4 for Africa, D2 for North and South America and B4, C4 and D4 for North Africa). Additionally, an increased cancer risks was detected for A3, A4 and D (sub) lineages (when compared with A1) in regions where they were common: A3 in East Asia; A4 in East Asia and North America and D in North and South/Central America where D lineages were also more frequent in ADC than SCC. Additionally, there was detected a variability in the cancer risk associated with different sub-lineages of HPV 16. Specifically, within lineage A, the A3 and A4 sub-lineages exhibited a greater risk of

cancer compared to A1. Similarly, among the BCD lineages, only D, and potentially only D2/D3, was linked to a higher likelihood of developing cancer.

In 2017 van der Weele *et al.* analysed the whole genomes of HPV 16 obtained from vaginal swabs with the aim of assessing whether particular variants were associated with integration in the context of persistent infection<sup>34</sup>. They found that out of the 17 HPV 16 A1–3 isolates, 47% were integrated into the host genome, while 48% of the A4 isolates were integrated. Additionally, there was no notable discrepancy in the prevalence of A1-3 and A4 sub-lineages among the groups with cervical intraepithelial CIN 1, CIN 2/3, and cervical cancer ( $p = 0.936$ ). A year later, the same group published a similar study with HPV 18 isolates obtained from self-collected cervical swabs<sup>35</sup>. They obtained the whole genome for 25 HPV 18 persistent infections and 26 transient HPV18 infections. According to their results, sub-lineage A3 was predominant in the population under study.

It has been proposed that some lineages may be associated with certain morphologies. Mirabello *et al.* (2015), looked at the different sublineages of HPV 16 and their association with SCC and adenocarcinomas. They found that variants A1/A2, D2 have a stronger risk of SCC, while D2/D3 and A4 sublineages were associated with glandular lesions<sup>31</sup>, B and C lineages were not associated with adenocarcinomas. Parraga *et al.* (2017) also found HPV 16 sublineages A1-3 were more prevalent in SCC and HPV 16 D, mainly D3, were increased in glandular cancer lesions<sup>36</sup>.

Differences in sub-lineage prevalence in different anatomical regions has also been investigated. In 2016, Parraga *et al.* analysed HPV 16 positive samples collected from different anatomical sites (cervix, vulva, vagina, penis, and anus) to determine differences in prevalence of the sub-lineages. However, the study

could not find significant differences in sub-lineage prevalence between the different anatomical regions<sup>36</sup>.

#### *1.6.2 Other genome variations and their association with cancer risk*

Recent analysis of the whole genome of human papillomavirus has revealed variations on specific locations linked to risk of progression to cancer and persistence of infection. In particular, the T350G polymorphism in the E6 gene has been associated with infection persistence and risk of progression to precancerous cervical lesions. The T350G substitution is non-lineage specific and corresponds to a single nucleotide change in the HPV 16 E6 gene. Cornet *et al.* (2013) looked at the type 16 variants focusing on E6 in France<sup>37</sup>. They found that both T and G at position 350 in the E6 gene are common in pre-cancer and cervical cancer in Northern Europe and that those infections containing 350T appeared to be more persistent vs 350G. Cornet *et al.* (2013) also looked at the incidence of variant lineages in Europe/Central Asia and South/Central America and found that distribution of HPV 16 350 SNP variants worldwide varies. European 350G isolates were less common in cervical cancer in East Asia and Europe/Central Asia), whereas the opposite was true in South/Central America<sup>38</sup>.

Mirabello *et al.* also studied the genetic conservation of the E7 gene and its link to carcinogenesis<sup>39</sup>. Authors evaluated genetic variants within HPV 16 and determined the associations between SNPs and risk of cervical pre-cancer/cancer. They discovered that the E7 gene from HPV 16 positive, but disease-free group, had an increased number of genetic variants compared to the E7 protein in pre-cancer/cancer samples. This demonstrates that the conservation of the 98 amino acids of E7, which act by disrupting the Rb function, is essential for HPV 16 carcinogenesis. This highlights the possibility of targeting this specific region for both etiological and therapeutic research.

## 1.7 Persistent HPV lesions and pathology classification

Most HPV infections are asymptomatic and will clear over time (90%), the other 10% will have an infection not cleared by the host (with an increased risk of developing pre-cancerous or cancer lesions <sup>40</sup>).

These lesions are termed depending on the region they appear and classified through histology and cytology study, by analysing the anatomy of tissues and organs and cells respectively.

The pre-invasive intraepithelial lesions are graded 1, 2 or 3 according to the level of dysplasia detected in the sample, with 3 being the most severe. Pre-cancerous lesions located in the cervix are designed as cervical intraepithelial neoplasia (CIN), For anal lesions, histology denomination is anal intraepithelial neoplasia (AIN). Similarly, preinvasive lesions of the penis, vagina and vulva are termed PeIN, VAIN and VIN, respectively. In the oropharynx, however, a pre-invasive/precursor lesion phase does not exist or has not been found <sup>41</sup>.

Intraepithelial neoplasia (both low grade and high grade) can clear without intervention; even with high-grade lesions and there is evidence that 47% of CIN3 lesions of lesions will regress naturally<sup>42</sup>. However, a component of persistent untreated lesions can progress to cancer, so in countries which have screening programmes, treatment of lesions is performed to prevent this. Given that lesions can regress without treatment, the ability to predict which lesions will progress vs regress would help in avoiding unnecessary treatment and associated morbidity<sup>42</sup>.

Cancers are also classified in distinct stages according to size and the extent of lesion-spread<sup>43</sup>. Classifications vary by anatomical site and world regions. The FIGO (International Federation of Gynaecology and Obstetrics) staging system is used most often for cancers of the female reproductive organs, including cervical cancer <sup>44</sup>. This classification has 4 levels: I, II, III and IV however I has 7 distinct

levels, stage II 4 levels, stage III 3 levels, and stage IV 2 levels. These stages reflect the size of the tumour, spread into lymph nodes, and spread into distant sites.

For other cancers, the staging system most used is the American Joint Committee on Cancer (AJCC) TNM system <sup>45</sup>, having 5 distinct levels according to size, lymph nodes affections and spread into distant sites: 0, I, II, III, and IV.

In addition to the cancer stage classification, lesions can also be classified attending to the type or morphology of cells involved in the lesion. There are several different types of cancer defined according to the nature of the affected cells:

- Squamous cell carcinoma (SCC): A type of cancer that originate from flat, skin-like cells that form the outer layer of the cervix, (known as the ectocervix). These types of cancer constitute the majority, around 70 and 80 out of every 100 cervical<sup>46</sup>. Almost all the oropharyngeal cancers are SCC while 9/10 anal cancers are SCC<sup>47</sup>.
- Adenocarcinoma (ADC): Adenocarcinoma is a form of cancer that begins in glandular cells responsible for producing mucus. Within the cervix, glandular cells are distributed along the inner lining of the channel connecting the cervix to the uterus (known as the endocervical canal). Although less frequent than squamous cell carcinoma, the incidence of adenocarcinoma has increased in recent years. More than 10 in every 100 cervical cancers (more than 10%) are adenocarcinomas<sup>48</sup>. In the anal regions, adenocarcinomas are very rare, reported in 5 to 10% of all anal cancers<sup>49</sup>.
- Adenosquamous carcinoma (ASC): Type of cancer that involves both squamous and glandular cancer cells. This variant of cervical cancer is relatively uncommon, accounting for only about 5% to 6% of all cervical cancer cases. Adenosquamous carcinoma is managed using similar treatment approaches to those utilized for squamous cell carcinoma of the cervix.

## 1.8 HPV detection – conventional tests

Like in other areas of microbiology diagnosis, there are several methods that can be used to detect HPV. The majority of commercially available tests rely on the amplification of HPV and the detection of its nucleic material (either DNA or mRNA). Other tests use a hybridization approach, where HPV is captured and results in fluorescence emission or a change in colour, but due to a higher in sensitivity, specificity, and robustness, most of the tests used these days are based on an amplification approach.

The majority of these test are based on nucleic acid amplification and not only detect the presence or absence of HPV but also provide information about the type or group (hr-HPV or lr-HPV) present. Most of the tests normally target a small region of L1 (mainly used for type differentiation) but some other target E6/E7 region, like Aptima (Hologic) and Xpert (Cepheid). Most service laboratories use commercial kits, however some labs (often but not always research) use in-house tests, developed by the laboratory, or translated from published research.

Types included in the commercial tests vary between kits. Some kits only identify high-risk types as one group and do not provide type specific information. Others provide type information for 16 and 18/45 and aggregate the rest of high-risk types in one group. Other tests provide HPV status for all types included in the kit. As an example, the Anyplex II HPV28 (Seegene) provides presence/absence information of 28 types including all established high-risk types. Moreover, tests also differ in the turnaround time and sample capacity. There are tests that can provide a result in less than an hour and others in approximately 3 hours, some of them also performing DNA extraction and genotyping in 6-8 hours. In terms of sample capacity, there are tests that can test 1 sample individually to several hundred. Choice of test will clearly depend on desired application, but high throughput testing is clearly attractive for population-based applications including screening.

There are a number of HPV tests including those used for cervical screening that are well established and evidenced in terms of their performance tests<sup>50</sup>. However, existing tests indicate presence or absence of the virus rather than activity or capacity for transformation. More specific tests that can delineate clinically significant from transient HPV infection is a key aspiration on the HPV scientific community. We know now that there are some variations in the HPV genome that have been associated with higher risk of cancer<sup>31</sup> or persistent infection<sup>51</sup>. Detecting these variations/mutations could result in better handling of the infection management of the disease. But the majority of the conventional tests won't be able to detect these variations in the HPV genome. Thanks to the development of new sequencing technologies, deep interrogation (identification of each base of the DNA) of the HPV genome or analysis of the integration regions of the HPV in the human genome are now possible and will be possible to identify these variations and use them as biomarkers.

## **1.9 Next generation sequencing and whole genome sequencing**

NGS is a technology based on massively parallel sequencing or “deep sequencing” of nucleic acid sequences. The target nucleic acid is fragmented in small pieces and each fragment is read thousands of times, providing a depth of information which can deliver accurate data at the nucleotide level. Additionally, NGS allows to sequence the entire genome of HPV or can be limited to specific areas of interest<sup>32,52-54</sup>.

### *1.9.1 NGS applications for other fields*

Due to the reduction of cost associated with NGS over the years, better understanding of the applications for microbiology diagnostics and the large investment and capital purchase performed by governments during the covid pandemic, service and diagnostic labs are increasingly adapting NGS as a diagnostic tool. NGS allows discovery of new viruses or bacteria, variants/lineages/serotypes, identification of organisms



difficult to culture<sup>55</sup>, mutations in the genome and detection of resistance to antimicrobials. It may also be used as a rapid response to outbreaks of pathogens<sup>56,57</sup>. For example, during the SARS-CoV-2 outbreak in the United Kingdom, sequencing technology was used to identify the entry route of the virus in Scotland<sup>58</sup>, different variants present in the community<sup>59</sup> and investigate outbreaks<sup>57</sup>.

All the recent advantages in next generation sequencing (NGS) have opened new perspectives in the field of clinical diagnostics by providing a better and personalised diagnosis and having a broader diagnostics repertoire.

### 1.9.2 NGS and HPV

While NGS has been used in HPV research for some time, its use in screening and clinical diagnosis is not yet widespread due to a lack of standardization and quality guidelines, as well as high costs and requirements for specialized laboratory infrastructure. However, the use of NGS could provide clinicians with additional information about HPV-driven lesions, improving diagnosis and management. This includes more sensitive detection of HPV, accurate quantification of HPV identification of HPV sub-lineages associated with increased risk<sup>31</sup>, and detection of circulating HPV DNA in oropharyngeal cancer cases<sup>60</sup>.

Unlike many molecular diagnostic assays/instruments, where it is possible to get a result straight from the machine, NGS requires the processing and analysis of large amounts of raw data using bioinformatic tools before genotypes can be assigned. Bioinformatics is the discipline that has developed methods and software tools to help interrogate and organize the data obtained from NGS instruments, including aligning the genome using a reference (mapping) or without (*de novo*) when sequence is unknown, organism identification, analysis of mutations, SNPs variations and phylogenetic analysis.

### *1.9.3 Next generation sequencing and HPV applications*

NGS has several potential research and diagnostic applications and is also seen as a promising method for HPV typing due to his high sensitivity even for types at low copy number within multiple infections, its potential ability to detect a broad spectrum of HPV types including novel types and its capacity to determine lineages and sub-lineages with accuracy. NGS alleviates many of the issues that are associated with PCR-based methods for detecting HPV, including false-positives due to cross-reactivity between different HPV types and false-negatives associated with low viral loads<sup>61,62</sup>. In addition, NGS can also be used to look at SNPs associated with higher risk of persistent infection or to identify viral integration in the human genome by looking at the whole genome of the virus.

There are multiple NGS protocols and tools that can be used for HPV sequencing, including short read, and long read sequencing, different library prep kits etc. But, as the majority of cancer samples tested for HPV are preserved as formalin-fixed paraffin embedded (FFPE), quality and quantity of the DNA available for testing could be impacted. The preservation process can have a negative impact on the sample DNA.

As it is a novel technology in the HPV field, there is a lack of consensus in the HPV diagnostic community about the protocols/approaches required to ensure a robust end to end process including interpretation of results. Consequently, it is necessary to explore the best approaches for validation and quality assurance procedures at each step of the NGS workflow would be of value.

## **1.10 Burden of diseases driven by HPV**

### *1.10.1 Worldwide*

As mentioned before, HPV infection is one of the most common infections in the world. However, the distribution and burden of disease varies depending on location and affluence of the countries. De Martel

*et al.* (2017) published an article where they looked at the worldwide burden of cancer attributable to HPV by site, country, and viral type. They found that 4.5% (AF) of all cancers worldwide (630,000 new cancer cases per year) are attributable to HPV, being 8.6% in women and 0.8% in men<sup>63</sup>. This value also varies between countries, from 2% in high-income countries (HIC) like Australia and the USA to >20% in India and sub-Saharan Africa.

The number of cases attributable to HPV varies depending on the anatomical site, described in Table 4. For cervical, anal, and vulvar cancers, the AF is high, 99%, 88% and 78% respectively. For penile, vaginal, and oropharyngeal cancers, AF is much lower with 50%, 34.9% and 30.8% respectively. Moreover, these values vary depending on world locations, Table 5. Cervix cancer accounts for 83% of HPV-associated cancer and 2/3 occurring in the low and middle-income countries (LMIC). However, oropharyngeal HPV-attributable cancer has a higher presence in the HIC countries, with 2.5 times high prevalence than LMIC. Data from 2017, for anal cancer, indicates that the estimated number of cases per year is 40000, while 35000 (88%) are attributed to HPV with similar distribution between males and females (17000 and 18000 respectively) and a higher incidence in those >50 years old than in the younger population (Table 2)

**Table 2. Number of all cancer cases attributable to HPV and corresponding attributable fraction (AF) for all cancers, cancer site(s) and sex.** Table prepared from data presented at de Martel et al. International Journal of Cancer, 2017<sup>63</sup>, data GLOBOCAN 2012.

Worldwide				Number attributable to HPV by gender - world		Number attributable to HPV by age group world		
	Number of cases	Number attributable to HPV	AF (%)	Male	Female	<50 years	50 - 69 years	70+ years
<b>Cervix</b>	530000	530000	99.0	0	5300000	250000	220000	58000
<b>Anus</b>	40000	35000	88.0	17000	18000	6600	17000	12000
<b>Vagina</b>	34000	8500	24.9	0	8500	2600	3400	2500
<b>Vulva</b>	15000	12000	78.0	0	12000	2500	5200	3900
<b>Penis</b>	26000	13000	50.0	13000	0	2700	5800	4400
<b>Oropharynx</b>	96000	29000	30.8	24000	5500	5400	18000	6000

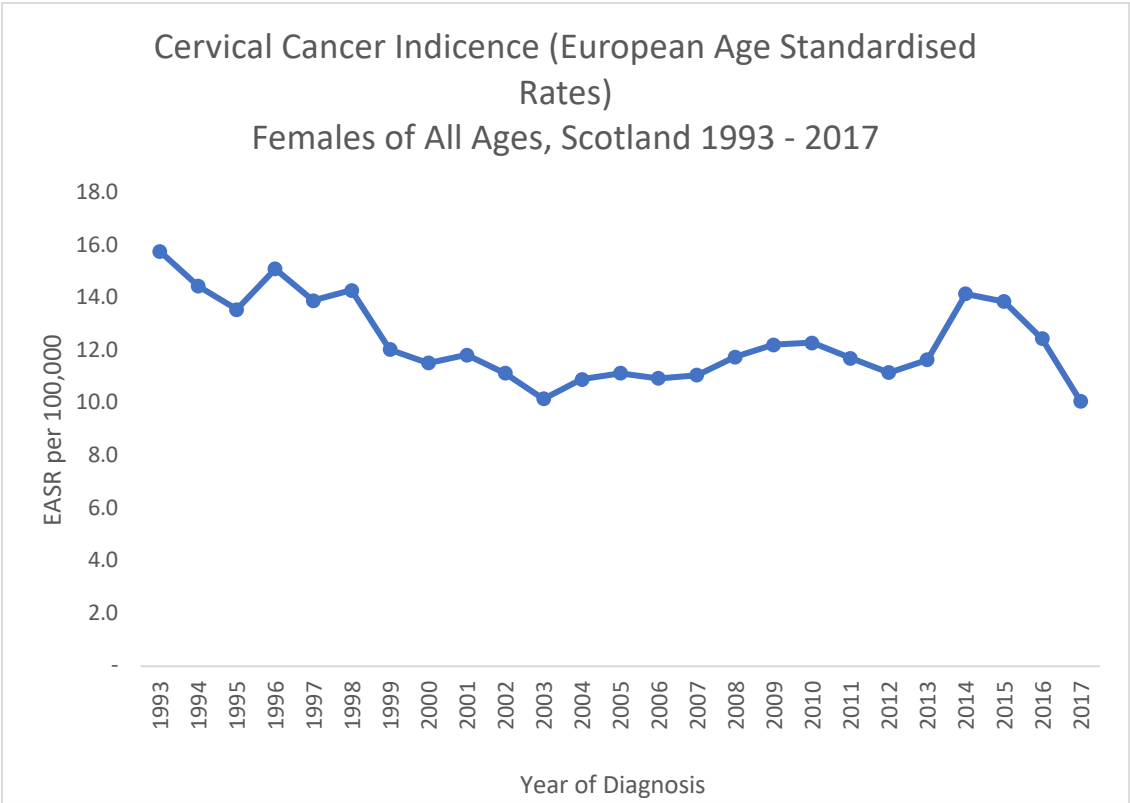
**Table 3. Number of different cancer cases attributable to HPV for various locations.** Table prepared from data presented at de Martel *et al.* International Journal of Cancer, 2017<sup>63</sup> (using data collected from GLOBOCAN 2012). Data for Scotland obtained from Information Service Division (ISD) Scotland<sup>64</sup> (Year 2017). Accessed 10/01/19. Numbers were rounded to two significant digits.

Region	Number of cases per year				
	Cervix	Head & Neck		Anus	
	Female	Female	Male	Female	Male
<b>Low-income countries</b>	370000	2100	8600	7600	10000
<b>High income countries</b>	160000	5500	22000	10000	6800
<b>Scotland</b>	275	360	910	90	50
<b>Europe</b>	58000	2800	11000	4200	2700
<b>North America</b>	14000	1900	7000	2700	1800
<b>Australia</b>	940	80	290	190	150

#### 1.10.2 Burden and epidemiology of disease in Scotland

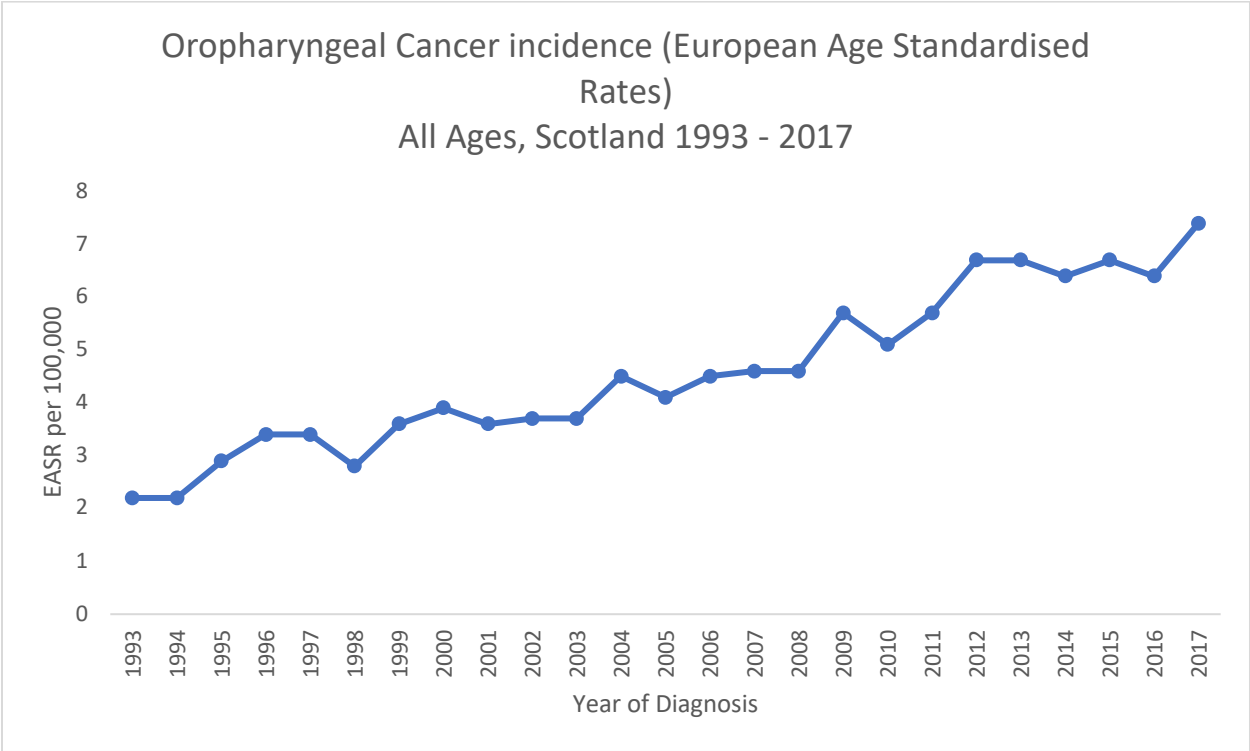
The Information Service Division (ISD) of National Services Scotland, part of NHS Scotland, has been collecting national data on cancer since the 1980s<sup>64</sup>. Figure 3 to Figure 5 show the incidence rate per 100,000 (EASR) of cervical, oropharyngeal, and anal cancers in Scotland from 1993 to 2017.

Figure 3 shows that from 2013, a decrease trend in cervical cancer incidence EASR (European age standardised rates per 100,000) was registered. Reducing from 14 cases per 100,000 (EASR) in 2014 to 10 cases per 100,000 (EASR) in 2017.



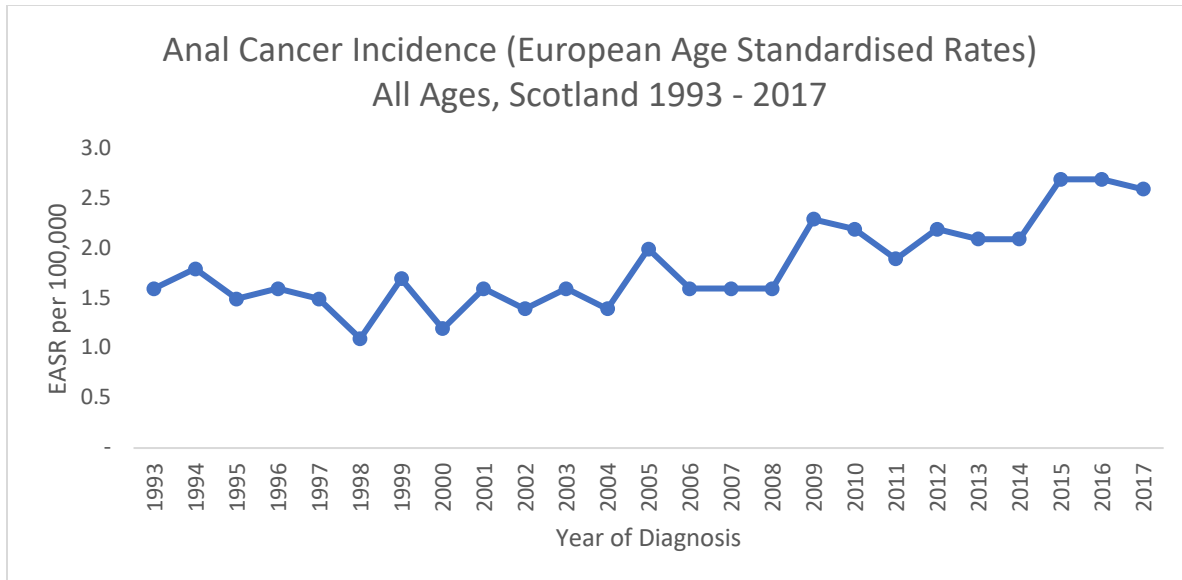
**Figure 3. Cervical cancer incidence in Scotland from 1993 to 2017.** Crude rate is calculated per 100,000 person-years at risk. Obtained from Information Service Division (ISD) Scotland. Accessed 10/01/19.

Comparatively, the European age-standardised incidence rate (EASR) of oropharyngeal cancer incidence per 100,000 has increased over the last 24 years (Figure 4), going from 2 cases per 100,000 (EASR) in 1993 to 7 per 100,000 (EASR) in 2017 for all ages and sex.



**Figure 4. Oropharyngeal cancer incidence (European Age-specific rates), Scotland: 1993 – 2017.** Obtained from Information Service Division (ISD) Scotland. Accessed 10/01/16.

An increase in incidence has also been registered for anal cancers, where it increased from 1.5 EASR in 1993 to almost 2.5 EASR in 2017.



**Figure 5. Anal cancer incidence (European Age-standardises rates), Scotland: 1993 – 2017.** Obtained from Information Service Division (ISD) Scotland. Accessed 10/01/19.

Although anal cancer incidence increased from 1993 to 2017, there is little epidemiological data on HPV type specific prevalence in high-grade anal lesions and anal cancer in the UK or Scotland. However, the prevalence of HPV in residual rectal swabs from asymptomatic men attending GUM clinic was assessed to create a baseline prior to the men who have sex with men (MSM) vaccine program<sup>65</sup>. According to the paper, out of all the swabs that were tested, 72.8% of them were found to be positive for HPV. Of these, 59.1% of samples tested positive for at least one high-risk type of HPV.

## **1.11 Disease management in Scotland: vaccination and screening**

### **1.11.1 Vaccination programs**

In 2008, Scotland became one of the first countries to implement a vaccination program which included schoolgirls between 11 to 13 years old. At the same time, a catch-up program was offered to girls up to age 18 (17 years and 364 days). The vaccination program initiated with the bivalent HPV vaccine (Cervarix,

GlaxoSmithKline) for the first four years which gives protection against high-risk types 16 and 18 and was provided as a 3 doses schedule. In 2012, the bivalent vaccine was replaced by the quadrivalent one, Gardasil (Merk & Co), a vaccine which included the same types as the Cervarix but also conferred protection against low-risk types 6 and 11. Two years later, in September 2014, the quadrivalent vaccine was offered to girls in a 2 doses program with at least 6 months between doses following the recommendation from the Joint Committee on Vaccination and Immunisation (JCVI).<sup>66</sup>

In July 2017, HPV vaccination was also offered to men who have sex with men (MSM) following the recommendation by JCVI. This decision was driven in part by the fact that while being at significantly increased risk of HPV associated disease, MSM would not gain the value of herd protection offered by what was a female only programme up to 2019. For the MSM programme the HPV vaccine was initially offered to men up to 45 years old within a 3 doses schedule in a 12-month period.

From 2019, the Scottish HPV vaccination program became gender neutral, as it was offered to schoolboys. Two years later, during the 2021/22 academic year, the quadrivalent vaccine was replaced by a nonavalent HPV vaccine (Gardasil-9), which offered protection against types 6, 11, 16, 18, 31, 33, 45, 52 and 58.

From 2023, the nonavalent vaccine will be administered routinely as 1 dose, following the latest recommendation of the JCVI (2022). The changes in vaccine type, targeted group, dosing since 2008 will very likely change the epidemiology of infection and disease in Scotland including and beyond the cervix and in both sexes. It will be important to monitor these changes to determine vaccine impact as well as implications for managing residual disease.



### 1.11.2 Vaccination rates

HPV vaccination had a high uptake among the schoolgirl cohort with a complete vaccination schedule exceeding 80% from 2014/15 to 2018/19 (Table 4). Data from 2016/17 indicated that 86% of girls from S1 were vaccinated<sup>67</sup>. In the 2021/22 academic year, the first dose coverage of HPV vaccine for S1 pupils increased to 73.5% from 52.1% in the previous year (due to covid driven school closures). Despite this bounce back, there is still a gender disparity, with female coverage at 77.5% and male coverage at 69.6%, representing a 7.9% difference. See full report in:

<https://publichealthscotland.scot/media/10789/2021-12-14-hpv-report.pdf>

**Table 4. Trend in coverage of first dose of HPV immunisation by the end of school years 2014/15 - 2021/22 in Scotland.** Pupils in S1 by sex. Data obtained from Public Health Scotland<sup>68</sup>.

Year	First dose coverage rate (%)		
	Both sexes	Female	Male
2014/15	..	89.0	..
2015/16	..	86.7	..
2016/17	..	85.6	..
2017/18	..	83.7	..
2018/19	..	85.1	..
2019/20	58.0	61.0	55.1
2020/21	52.1	54.7	49.6
2021/22	73.5	77.5	69.6

### 1.12 Epidemiology of HPV in Scotland

Before the introduction of the HPV vaccine in the national program, different studies were performed which assessed the prevalence of HPV in the general population and in those affected by cervical cancer.

This was performed to determine a baseline to from which to compare changes brought about through implementation of the HPV vaccine. In 2010, Cuschieri *et al.* analysed the prevalence of HPV in cervical cancer biopsies collected between 2004 to 2008<sup>68</sup>. Samples were collected in all main pathology centres in Scotland. A high prevalence of HPV high-risk types 16 and 18 were present in 82% of the HPV positive cervical biopsies analysed. The next most frequently detected types in the Scottish invasive cervical cancer (ICC) cases were, in order of prevalence, HPV 45, 33 and 31. This helped to determine type specific burden in invasive disease.

A year later, O'Leary *et al.* assessed the HPV type specific prevalence of unvaccinated 11-18 years in urine samples to inform effectiveness studies for the new HPV immunisation program in Scotland<sup>69</sup>. The authors found a higher prevalence of HPV in women between 14 to 18 years old than 11-14 years, 15.2% and 1.1% respectively. Furthermore, 6.5% of the 14-18 years group were infected with HPV 16 and 18. The study provided a baseline of the prevalence in the unvaccinated young population.

After the start of the HPV national vaccination program, different Scottish studies looked at the prevalence of HPV in different population to assess if the vaccination was having any impact in HPV prevalence or lesions<sup>70-79</sup>.

Cameron *et al.* (2016) published the prevalence of HPV types in girls and the catch-up (girls up to age 18) population after the introduction of the vaccination program between years 2009 and 2013<sup>71</sup>. They tested 4,679 samples, finding that three doses of bivalent vaccine were associated with a significant reduction in prevalence of HPV 16 and 18 from 29.8% to 13.6%. Significant reduction of prevalence was only observed after the 2<sup>nd</sup> and 3<sup>rd</sup> doses.

In 2017, Kavanagh *et al.* published the changes in the prevalence of HPV in the Scottish vaccinated population and the effectiveness of the vaccine during a 7-year period<sup>76</sup>. This analysis presented the first data from the routinely vaccinated cohort. In this study a reduction of the HPV types 16 and 18 was observed as well as a high degree of cross protection against types 31, 33 and 45. Data collected showed a reduction of types 16 and 18 from 30.0% in the pre-vaccinated cohort to a 4.5% in the vaccinated cohort. Additionally, vaccine effectiveness was calculated, giving an 89.1% protection against HPV infection in those vaccinated at age 12 – 13 years old. A significant vaccine effectiveness for cross protective types 31, 33 and 45 of 93.8%, 79.1% and 82.6% respectively was registered. Furthermore, the publications by Cuschieri *et al.*<sup>70</sup> and Kavanagh *et al.*<sup>76</sup> could not detect any viral type of replacement after the introduction of the vaccine in Scotland.

Additional studies have looked at HPV prevalence in sites in non-cervical sites in Scotland. The HOPSCOTCH study looked at the prevalence of HPV in the oral cavity, Conway *et al.* 2016<sup>80</sup>. A total of 402 individuals participated in the study providing 10ml of oral rinse in saline which were analysed for HPV genotype. HPV was detected an overall prevalence of 5.5% of the participants. According to age group, prevalence in 20 – 30 years old group was 7%, and 11% for the 50-60 years old group. Moreover, the higher prevalence noted was among the 26 – 39 years old group with 10.4%. No significant differences were detected between the sexes, with 5.2% and 5.8% for male and female, respectively.

A recent publication from Schache *et al.*, 2022, looked into the HPV-related oropharynx cancer in the UK<sup>81</sup>. They found that 45.4% of the samples collected in Edinburgh (serving Lothian, Dumfries and Galloway, Fife and Borders) were HPV positive, while the UK HPV positivity was 51.8%.

In terms of anal cancer, no studies in Scotland or including Scottish samples have looked into HPV prevalence in anal lesions or cancer.

### 1.13 HPV screening and surveillance in Scotland

Scotland benefits from primary and secondary prevention of cervical infection (and associated disease) through immunization and screening respectively. An HPV epidemiology and surveillance strategy was conceived in 2008 by the main public health agency in Scotland in order to determine the impact of HPV vaccine on HPV infection and associated disease<sup>66</sup>.

Cervical screening was first implemented in Scotland during the 1960s, but it was not established as a population-based program until 1988. Despite the availability of tests to large numbers of women, the service was not officially introduced as a screening program until that time. The national cervical screening program was introduced with the aim to detect, in initial stages, abnormalities or changes in cervical cells which if left untreated could develop into cervical cancer. Primary screening has been based on HPV testing since 2020 with cytology as a triage<sup>82</sup>; depending on the result, women would be referred to colposcopy, recalled in 3-6 months (women with low grade smears) or in 5 years. Now, the program is available to women between 25-65 years to attend every 3 years for women from 25-49 years old and every 5 years to women from age 50-64 plus 364 days.

In 2018, a longitudinal HPV Surveillance program implemented on residual liquid-based cytology sample (LBC) from women who attended their first cervical screening (20 years old).

In terms of surveillance, different programs have been performed in Scotland over the years.

- **2013 – 2015:** Oral prevalence in Scotland. HOPSCOTCH study. A study designed to investigate the prevalence, incidence, and persistence of oral HPV in Scotland via dental settings<sup>80</sup>.

- **2015/2016** Residual LBC were obtained from the first group of women vaccinated during the school years that attended their first cytology screening.
- **October 2016 – February 2017:** HPV prevalence on Men who have sex with men (MSM). Rectal swabs included in the study were collected and taken from men who attended for an asymptomatic sexual health screen or for treatment of a presumed STI.
- **2017-2018:** HPV prevalence on cervical CIN and cancer biopsies. Samples were collected in the main pathology centres in Scotland.

#### **1.14 Role of the Scottish Human papillomavirus Reference Laboratory**

The Scottish Human Papillomavirus Reference Laboratory (SHPVRL) was created in 2008, with the aim of identifying HPV for the purpose of aiding individual patient management and providing comprehensive epidemiological data for health protection initiatives. The laboratory is based at the Royal Infirmary of Edinburgh, and it is commissioned by National Services Division (NSD) and Public Health Scotland (PHS) to provide a service for Scotland.

The laboratory's main functions include the specialist diagnostic and advisory services for specific clinical cases where HPV testing is relevant including cervical and oropharyngeal cases. It is also involved in data monitoring, audit, and service developments related to the Scottish Cervical Screening program. A further key remit of SHPVRL is to deliver testing services to support epidemiology and surveillance of HPV and HPV associated disease in the Scottish population, particularly given immunisation.

The laboratory also provides guidance and materials related to quality assurance and assessment of HPV testing and is committed to a research and development program collaboration with NHS partners,

academia, and industry. It also consolidates HPV samples in a sample archive to support HPV-based research, test development and teaching (located at the HPV Research Group, University of Edinburgh).

## **1.2 Aims of the thesis**

Despite the studies looking at the HPV in cervical and oropharyngeal cancer in Scotland, there is lack of information in terms of the preventable fraction of HPV infections in these cancers by the HPV vaccine. There are multiple explanations, including the lower incidence of the non-cervical HPV-driven cancers in the Scottish population and the lack of pre-invasive stage (or not yet understood) in the case of oropharynx.

Incidence of anal cancer cases has increased worldwide<sup>83</sup> and in Scotland has increased +120% from 1976 to 2017<sup>84</sup>. Despite this, HPV prevalence in anal cancer in Scotland is unknown and even with the high HPV vaccine uptake registered in Scotland, it will take some time to see the full effect due to the age profile of anal cases.

Additionally, the role of HPV status and viral load has not been widely investigated in anal cancer, while in other HPV-driven cancers it has been demonstrated that HPV status and viral load can play an important role in overall survival. Also, even if some HPV 16 sub-lineages have been associated with higher risk of cervical cancer, no information is available for anal cancer.

Any reference laboratory that provides a specialist service it should be committed to evaluating and optimising new technologies that can support their clinical research and epidemiological remits optimally and be reactive to changing disease burden adapting new technologies and approaches. Thus, it is necessary to determine the best use of NGS for HPV diagnosis in a specialist service.

Trying to answer these questions, the main aim of the thesis is to perform a detailed molecular interrogation of HPV in anal cancer to determine and inform the impact of prevention and management strategies. To address this, this work intends:

- 1 To provide a contemporary description of HPV types in the most common invasive lesions in Scotland (including cervical, oropharyngeal and anal cancer).
- 2 To set up an HPV NGS system and pipeline within a reference laboratory context.
- 3 To address knowledge gaps on the status, diversity, and implications of HPV in anal cancer by:
  - Performing HPV type specific assessment in a well characterised series of anal cancer in Scotland.
  - Mapping HPV lineages and sub-lineages in both anal cancer asymptomatic cohorts.
  - Evaluating the association of qualitative HPV status on clinical outcomes
  - Evaluating the association of HPV viral load on clinical outcomes
  - Evaluating the implications of lineages/sub-lineages profile with current disease status and future survival outcomes.
- 4 To identify the best NGS protocols to implement into an HPV diagnosis laboratory.

## 2. General Materials and Methods

### 2.1 Materials

**Table 5. Consumables and Manufacturer**

Consumables	Manufacturer
Qiagen DNA Mini Kit	Cat#: 51304, Qiagen, Hilden, Germany
Anyplex II HPV28 Detection	Cat#: HP7S00X, Seegene, Seoul, Korea
Seegene Universal extraction system	Cat#: 744300.4.UC384, Seegene
Seegene lysis buffer	Cat#: unknown, Seegene, Seoul, Korea
TruSeq DNA Nano	Cat#: 20015964, Illumina, San Diego, USA
TruSeq DNA Single Indexes Set A (12 Indexes, 24 Samples)	Cat#: 20015960, Illumina, San Diego, USA
TruSeq DNA Single Indexes Set B (12 Indexes, 24 Samples)	Cat#: 20015961, Illumina, San Diego, USA
Illumina® DNA Prep, (M) Tagmentation (24 Samples)	Cat#: 20018704, Illumina, San Diego, USA
Nextera™ DNA CD Indexes (96 Indexes, 96 Samples)	Cat#: 20018708, Illumina, San Diego, USA
MiSeq Reagent Kit v2 (500 cycles)	Cat#: 20018708, Illumina, San Diego, USA
Qubit™ dsDNA High Sensitivity	Cat#: Q32854 ThermoFisher, Waltham, USA
Qubit Broad Range Assay Kit	Cat#: Q32853 ThermoFisher, Waltham, USA
Qiagen PCR Multiplex kit	Cat#: 206143 Qiagen, Hilden, Germany
HPV 16 primers	Eurogentec, Seraing, Belgium
EcoRI in 1x NEB Cutsmart buffer	Cat#: R0101S New England Biolabs (NEB)



HindIII in 1x NEB Cutsmart buffer	Cat#: R0104S New England Biolabs (NEB)
Tapestation High Sensitivity DNA 1000 Reagent	Cat#: 5067-5585, Agilent, Santa Clara, USA
Tapestation High Sensitivity DNA 1000 Screen Tape	Cat#: 5067-5584, Agilent, Santa Clara, USA
Tapestation DNA 1000 Reagent	Cat#: 5067-5583, Agilent, Santa Clara, USA
Tapestation DNA 1000 Screen Tape	Cat#: 5067-5582, Agilent, Santa Clara, USA
Optiplex HPV Genotyping Kit	Cat#: IN0601, DiameX, Heidelberg, Germany
QIAamp MinElute	Cat#: 28604, Qiagen, Hilden, Germany
QIAamp DNA FFPE Tissue Kit	Cat#: 56404, Qiagen, Hilden, Germany
SYBR Green PCR MasterMix	Cat#: 4344463 ThermoFisher, Waltham, USA
PCR grade water	Cat#: 733-2573, VWR, Radnor, USA

## 2.2 Equipment

**Table 6. Equipment and Manufacturer**

Equipment	Manufacturer
CFX96 Touch Real-Time PCR Detection System	Bio-Rad, Hercules, USA
Microlab Nimbus	Hamilton Robot, Reno, USA
MiSeq	Illumina, San Diego, USA
Microtome – HistoCore MULTICUT	Leica, Wetzlar, Germany
GeneAmp PCR System 9700	ThermoFisher, Waltham, USA
Water bath	GAL,

Tapestation 4200	Agilent, Santa Clara, USA
QX200 Droplet Generator	Bio-Rad, Hercules, USA
QX200 Droplet Reader	Bio-Rad, Hercules, USA
Quantasoft	Bio-Rad, Hercules, USA
Luminex LX200	Luminex, Austin, USA

## 2.3 Nucleic acid extraction and HPV genotyping

### 2.3.1 Cervical and oropharyngeal cancer samples

Please note that no cervical or oropharyngeal cancer samples were extracted as part of this doctoral thesis. Instead, the results described in chapter 3 relate only to the assessment of the data. However, for a better contextualisation, extraction and genotyping protocols are described below.

#### 2.3.1.1 Cervical cancer samples extraction

The paraffin blocks received at the Scottish HPV Reference Lab (SHPVRL) as part of the national immunisation surveillance were obtained from the pathology laboratories and sectioned (10 µm) using a Leica microtome. Nucleic acid was extracted using the QIAamp DNA Mini Kit (Qiagen, Heidelberg, Germany) using an optimised protocol developed by Steinau *et al.* (2011)<sup>85</sup> for HPV recovery from FFPE samples. A total of 200 µl of ATL was added to each sample, followed by a 20-minute incubation at 120°C. Tubes were vortex for 10 seconds and quick spin. 20µl of Proteinase K was added to the tube followed by a 10 second vortex and spin down. Samples were then incubated at 64°C in a water bath overnight. The next morning, 200µl of AL was added to each sample and incubated at 70°C in a hot block for 10 min. 200µl of 100% ethanol was then added to each tube and incubated for 5 minutes. All lysate was then transferred into a Qiamap spin column and centrifuged at 8000rpm for 1 min. Columns were then washed

with 500 µl of AW1 and AW2 and centrifuged at 8000 and 13000rpm respectively. 200 µl of elution buffer was then added to each column, incubated for 1 minute at room temperature and centrifugated at 8000rpm for 1 minute. Eluate was then stored at -80°C.

#### *2.3.1.2 Clinical oropharyngeal samples extraction*

Clinical oropharyngeal sample sections (10 µm) were received at the SHPVRL accompanied with a request form. Extraction was performed using the process described in the above section (2.3.1.1).

#### *2.3.1.3 HPV detection of cervical and oropharyngeal samples*

The test selected by the reference lab to genotype the samples was the Optiplex HPV Genotyping Kit (Diamex, Heidelberg, Germany). It is a Polymerase Chain Reaction (PCR)-amplified method for the detection of 24 of the most common HPV types: 14 high-risk types (16, 18, 31, 33, 35, 39, 45, 52, 56, 58, 59, 68) “carcinogenic or probably carcinogenic”, 6 putative types (26, 53, and 66, 70, 73 and 82) “possibly carcinogenic” and 6 “low-risk” types: (6, 11, 42, 43, 44) <sup>86</sup>. The HPV genotyping is based on polystyrene beads dyed with different fluorophores detected and quantified by a Luminex analyser (LX-200).

A median fluorescence intensity (MFI) relative to each bead-&-probe was calculated for every sample and the MFI recorded. An MFI of equal to or greater than the cut-off value indicates a positive HPV result whereas an MFI measurement less than the cut-off indicated the absence of HPV DNA sequences or HPV DNA levels below the level of detection limit of the assay.

#### *2.3.2 Anal lesions:*

A 10 µm section was obtained from each block using a microtome (Leica HistoCore MULTICUT, Germany). A deparaffinization step was performed prior to the extraction with 300 µl of Seegene lysis buffer to each 10 µm sections and overnight incubation at 64 °C degrees.

#### *2.2.3.1 Anal cancer - DNA extraction*

The nucleic acid extraction was performed using the Microlab Nimbus (Hamilton, Reno, USA) automated system and the Seegene Universal Lysis buffer. This instrument allows extraction and mastermix preparation for 42 samples thanks to its robotic pipette. However, FFPE samples require a deparaffinization step before addition into the machine. This consisted in adding 300 µl of Seegene lysis buffer to each 10 mm sections and incubated overnight at 64 °C degrees in a water bath. The following day, tubes were collected from the water bath and centrifuged for 10 seconds at full speed (13000rpm). Supernatant was transferred to 1.5 ml tubes. Tubes with the paraffin pellet were discarded.

The automated extraction was performed using Seegene's Nimbus protocol prepared for the instrument. DNA extraction was based on magnetic beads which join to the DNA and through several buffer washes the cellular debris are washed away. Final elution volume obtained was 100 µl.

#### *2.2.3.2 Anal cancer - genotyping*

HPV detection was performed using the Seegene Anyplex II assay. This assay detects 28 HPV types, identifying them as positive, negative, or invalid. It also provides semi-quantitative information about the number of copies of the L1 gene of the virus present in the sample. This is calculated based in the amplification detection cycles; at melting PCR cycles 30 (+++), 40 (++) and 50 (+). Types detected by the assay are: 12 high-risk carcinogenic types (HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59), probable carcinogenic types (HPV type 68), and some possibly carcinogenic types (HPV types 26, 53, 66,

70, 73, and 82) and low-risk types (HPV types 6, 11, 40, 42, 43, 54, 44 and 61) as defined by the International Agency for Research on Cancer<sup>1</sup>.

Once the extraction was completed, the Nimbus instrument prepared automatically the mastermix using the Anyplex II reagents present in the kit. 5 ml of EM1, 5 ml of RNase Free water and 5 ml of Mastermix A and 5 ml of Mastermix B.

A total of 5 µl of the sample nucleic acid was added to the 15 µl of the mastermix to generate a final volume of 20 µl. Wells were closed with optical caps. Plate was transferred to the CFX96 thermocycler (Bio-Rad, USA) for the PCR detection with the cycles described below (Table 7. **Seegene Anyplex 28 II Melting PCR cycles.**) taking a total time of 3:35 hours.

**Table 7. Seegene Anyplex 28 II Melting PCR cycles.**

Step	Temperature	Duration	No of cycles
1	50°C	4 min	
2	95°C	15 min	
3	95°C	30 sec	30
4	60°C	1 min	
5	72°C	30 sec	
6	GOTO 3, 29 more times		
7	55°C	30 sec	
8	Melting curve 55°C ~ 85°C (5s / 0.5°C)		
9	95°C	30 sec	10
10	60°C	1 min	
11	72°C	30 sec	
12	GOTO 9, 9 more times		
13	55°C	30 sec	
14	Melting curve 55°C ~ 85°C (5s / 0.5°C)		
15	95°C	30 sec	10
16	60°C	1 min	

17	72°C	30 sec	
18	GOTO 15, 9 more times		
19	55°C	30 sec	
20	Melting curve 55°C ~ 85°C (5s / 0.5°C)		
Note: Plate Read at steps 8, 14 and 20. Fluorescence is detected at Melting			

Once the run was finished, raw-data was exported to the Seegene auto-interpretation software. This software merged the PCR data and generated an interpretation document with results for each sample. Invalid samples were repeated with a new section and diluted 1:10 before adding them into the Nimbus for extraction.

### Anyplex™ II HPV28 Detection (96 plate)

Sample No	Patient Id	Well	Name	Type	FAM		HEX		Texas Red			Cy5		Quasar 705			Cy5	Auto Interpretation	Comment		
					66	45	58	51	59	16	33	39	52	35	18	56				68	31
		A01	HPV121	SAMPLE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	++	16(++)	
		A07	HPV121		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		B01	HPV122	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	16(+++)	
		B07	HPV122		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		C01	HPV123	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	-	
		C07	HPV123		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		D01	HPV124	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	16(++)	
		D07	HPV124		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		E01	HPV125	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	16(++)	
		E07	HPV125		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		F01	HPV126	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	16(++)	
		F07	HPV126		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		G01	HPV127	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	16(+++)	
		G07	HPV127		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		H01	HPV128	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	-	
		H07	HPV128		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		
		A02	HPV129	SAMPLE	66	45	58	51	59	16	33	39	52	35	18	56	68	31	IC	Invalid	
		A08	HPV129		26	69	73	42	82	53	43	54	70	61	6	44	40	11	IC		

Figure 6. Example of Anyplex II HPV28 report. Report is automatically produced by the Seegene interpretation software.

#### 2.3.3. Residual rectal swabs

In chapter 6 as a control group, residual rectal swabs obtained from asymptomatic men attending sexual health clinics were sequenced for HPV 16 sub-lineage identification. These samples were part of a published study (Cameron *et al.* 2019) where they looked at the HPV prevalence in residual rectal swabs from MSM attending health clinics<sup>65</sup>.

*2.12.4 Nucleic acid – residual rectal swabs*

Nucleic acid from these samples were previously extracted for the original study using a combination of the Qiagen MDx Robot (Qiagen, Germany) and the Seegene Universal extraction kit. Nucleic acid obtained from both extraction systems was stored at -80°C.

**2.3.4 DNA extraction methods used in the nucleic acid comparison for the identification of the best method for NGS downstream**

There exist different DNA extraction methods, each one with different characteristics, including the purification technology, overnight incubation, final elution volume and use of robot. As described in Table 8, these characteristics were different depending on the method. Manufacturer information of each of these kit is described in Table 5.

**Table 8. Characteristics of the 3 nucleic acid methods examined in chapter 5.**

Characteristics	Qiagen GeneRead FFPE Kit	Qiagen DNA Mini Kit	Seegene Universal extraction system
<b>Purification technology</b>	Silica gel membrane technology	Silica gel membrane technology	Magnetic beads
<b>Overnight incubation?</b>	No	Yes	Yes
<b>Elution volume</b>	20 – 40 µl	200 µl	100 µl
<b>Automation?</b>	No	No	Yes

#### 2.3.4.1 Extraction comparison – FFPE Microtomy

Three 10 µm sections were obtained from each block using a microtome (Leica HistoCore MULTICUT, Germany) following the SHPVRL Standard operating protocol for microtome sectioning.

Deep cleaning of the instrument before and between samples was carried out using 10% bleach and a new microtome blade was used for each sample. Moreover, a blank negative and a positive block were sectioned every day to reduce and identify the risk of contamination between samples. Sections were allocated in 2ml Sarstedt tubes.

#### 2.3.4.2 Extraction comparison – Qiagen DNA Mini kit

Extraction was performed using the Qiagen DNA Mini kit, protocol described in section 2.3.1.1.

#### 2.3.4.3 Extraction comparison – GeneRead DNA FFPE kit

The GeneRead DNA FFPE Kit is designed to purify of high yields of DNA from small amounts of FFPE tissue sections. Additionally, the procedure includes the removal of deaminated cytosine to prevent inaccurate results in DNA sequencing. Each 10µm section was placed in a 1.5 microcentrifuge tube. 160µl Deparaffinization Solution was added to each tube and incubated at 56°C for 3 min. A mix of 55 µl RNase-free water, 25 µl Buffer FTB, and 20 µl proteinase K was added to each tube followed by a 1-hour incubation at 56°C followed by 1-hour 90°C incubation. The lower clear phase was then transferred into a new labelled microcentrifuge tube. 35 µl of Uracil-N-glycosylase (UNG) was then added to the sample and incubated at 50°C for one hour in a heating block. 2 µl RNase A (100 mg/ml) was added to the tube followed by an incubation of 2 minutes at room temperature. Then, 250 µl of AL buffer was added to each sample, followed by 250 µl for 100% Ethanol. Lysate was then added to a QIAamp MinElute column and centrifuged at maximum speed for 1 minute. Columns were then washed with 500 µl of AW1 and AW2



followed by another wash with 250 µl of 100% Ethanol. Finally, 40 µl of ATE buffer was then added to the QIAamp columns for elution. Elute was stored at -80°C.

#### *2.3.4.4 Extraction comparison – Seegene Automated extraction*

This extraction protocol was based on the Nimbus Microlab platform (Hamilton, Reno, USA) which automatically extracts the NA using magnetic beads. However, due to the presence of paraffin, this protocol required an overnight incubation at 65°C with Seegene lysis buffer.

Each FFPE sample was inserted in a 2ml microcentrifuge tube. 300 µl of Seegene Universal Lysis buffer was added to each of the FFPE sections, followed by a 5-second vortex and pulse centrifuge. Tubes were then incubated at 65°C overnight. The next day, tubes were centrifuged, and supernatant was transferred into a 1.5 ml microcentrifuge. Tubes containing the supernatant were transferred into the Hamilton Nimbus and were extracted using the Seegene automated extraction protocol. The Nimbus extraction process took approximately 90 minutes to extract 42 samples, obtaining an elution volume of 100 µl.

#### *2.3.4.5 Extraction comparison - Quantification of DNA*

##### *2.3.4.5.1 Qubit*

Sample DNA obtained from the 3 different extraction methods were quantified using the Qubit (ThermoFisher Scientific, Waltham, MA, USA) and dsDNA BR Assay. Process was performed according to the manufacture's protocol. 5 µl of DNA sample was used for each test into 195 µl of buffer.

##### *2.3.4.5.2 qPCR for the determination of HPV DNA present in the FFPE samples. Chapter 5.*

To determine the amount of HPV 16 nucleic acid present in each sample, a quantitative polymerase chain reaction (qPCR) test was used. qPCR was used as a tool to compare the output of extraction efficiency.

This qPCR was designed using PK02718\_PaVE sequence as a reference (HPV 16) and targeting the L1 gene sequence, resulting in an amplicon of 192bp (F: 6937 Stop 6957, R: 7127 Stop 7106). Primers and the sequence are described in Table 9. This assay was designed to obtain a long amplicon to be able to determine if exist DNA fragmentation in FFPE samples and if there are differences between methods.

**Table 9. Name and sequence of HPV 16 primers and probe.**

Oligonucleotides	Sequence (5' -> 3')
HPV 16 primer Forward (22bp)	CTCCAGCACCTAAAGAAGATCC
HPV 16 primer Reverse (22bp)	TTGTAGAGGTAGATGAGGTGGT
HPV 16 probe (24bp)	/56-FAM/ACAAGCAGG/ZEN/ATTGAAGGCCAAACC/3IABkFQ/

PCR reaction was carried out in a total volume of 20µl including 0.2 uM of primers, 0.1 uM of probe and 5µl of sample. Mastermix used was Qiagen Multiplex PCR Kit.

PCR was carried out under the following conditions: 95°C for 15 min, followed by 40 cycles of 30 secs at 95°C, 45 secs at 52°C and 45 secs at 72°C, using the CFX96 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, USA). Interpretation of the qPCR was performed using the CFX software. PCR reaction had a total volume of 25µl volume (21µl of mastermix and 4µl of template DNA). Mastermix contains 12.5µl SYBR Green PCR MasterMix (containing AmpliTaq DNA polymerase), 3µl of MgCl<sub>2</sub> (3µM), 0.75µl (10µM) of each primer and 1µl PCR grade water (VWR).

## **2.4 Governance**

### *2.4.1 Governance – Annotation of cervical and oropharyngeal cancer (data only)*

HPV typing information on cervical cancers was obtained as part of a National HPV Surveillance programme in Scotland, which is delivered under the auspice of Public Health Scotland (PHS) (formerly Health Protection Scotland). The SHPVRL is commissioned by National Services division to deliver HPV testing services to support the surveillance programme. All tissue samples were taken as per routine standard of care as a consequence of the management of cervical disease and genotyped centrally at the SHPVRL using the technology described in section 2.3. HPV Surveillance is supported through system of linkages permissions that relate to the national surveillance program, these include Caldicott guardian permissions, the Public Benefit and Privacy Panel for Health and Social Care (1617-0175) and a data protection impact assessment for which the SHPVRL is included as a partner : “The collection and use of data for the surveillance of the impact of HPV immunization on the incidence and mortality from cervical cancer and its precursors” As part of the system of linkage/data flow, no personal identifying information was received at the SHPVRL as a consequence of the surveillance exercise

HPV prevalence Data on oropharyngeal cancer was generated as part of the SHPVRL routine and national remit to provide identification and typing of HPV in order to guide the clinical management of individual patients and to supply detailed epidemiological information for health protection purposes as described in the service level agreement. SHPVRL receive OPC samples for HPV testing and as part of chapter 3 aggregate data (only) on positivity are presented, with no fields presented that are < n=10.

### *2.4.2 Governance – Collation and annotation of anal cancer cases (including samples)*

The project involved the use of archived anal cancer samples taken between 2009 and 2018 as part of standard of care for the management of patients with anal disease. Samples were archived as formalin fixed paraffin embedded tissue blocks at the Western General Hospital in Edinburgh. Use of samples for the present project was approved by the Southeast of Scotland National Research for Scotland (NRS Bioresource) Bioresource (REC reference Ref 20/ES/0061- SR 1283). Favourable ethical opinion to conduct the research was also provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482.

#### *2.4.3 Residual rectal swabs*

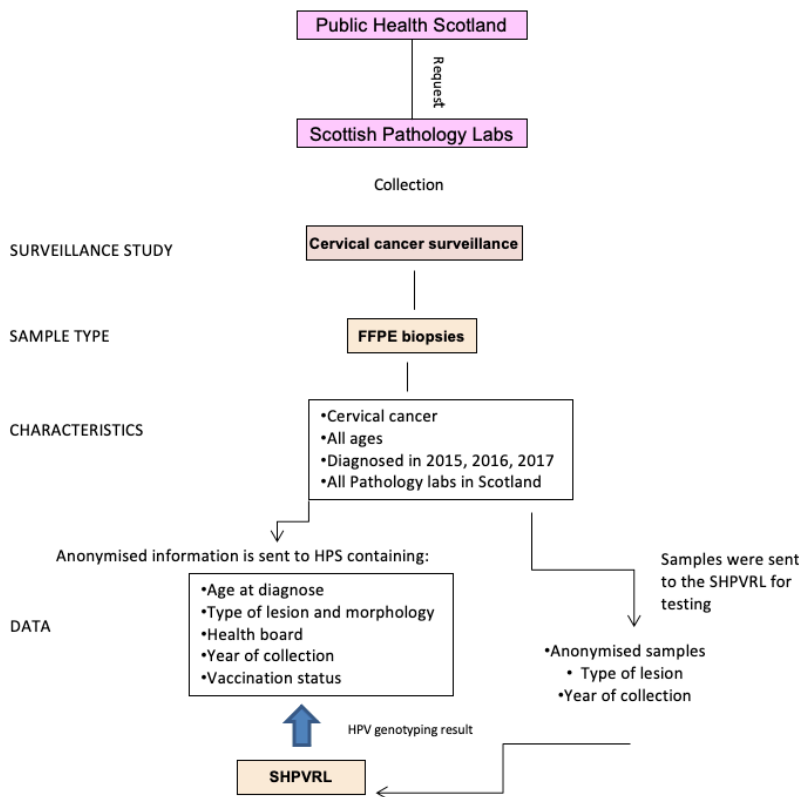
Residual rectal swabs were collected as a strand of national surveillance and obtained through permissions by the NRS Bioresource (REC reference Ref 20/ES/0061- SR 1283). Nucleic acid samples from these samples previously genotyped for HPV were archived as nucleic acid at the SHPVRL. Use of samples for the present project was approved by the Southeast NRS (REC reference Ref 20/ES/0061 application reference SR1364). Favourable ethical opinion to conduct the research was provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482.

## **2.5 Information capture for cervical and oropharyngeal cancer data**

### *2.5.1 Cervical cancer sample collection through National Surveillance*

As part of national surveillance study, each pathology laboratory across Scotland was requested to collate a case list of cervical cancers, irrespective of age, collected in 2015, 2016 and 2017. Laboratories were asked to identify a target number of samples related to the size of the NHS health board – e.g. Greater Glasgow and Clyde, which serves the largest population, were requested to send the greatest number of

cervical lesions. Pathology labs sent details of the identified cases with a surveillance study number, lesion morphology or histology and collection year to Information Services Division (part of PHS) where linkage to demographic information and vaccination status was performed. The pathology lab sent the formal fixed paraffin embedded (FFPE) biopsies to SHPVRL for HPV genotyping with the surveillance study number, morphology, and year of collection only (no patient identifying information was received). HPV results generated at SHPVRL were then sent to PHS with a study number which acted as a common bridge/link to demographic details captured by Information Services Division (ISD). Figure 7 describes the process of both surveillance programs involving high-grade and invasive cervical lesions.



**Figure 7. Diagram describing both Scottish cervical biopsy surveillance studies.**

A total of a total of 649 cervical invasive lesion samples were sent to the SHPVRL for HPV genotyping from the 3-collection years (2015, 2016 and 2017). Histological findings were classified according to the British

Association for Cytopathology and NHS cervical screening program criteria<sup>87-89</sup>. Cervical cancer and samples were classified in 3 different groups attending to their histology identification: Squamous cell carcinoma (SCC), adeno squamous carcinoma (ASC) and adenocarcinoma (ADC).

### 2.5.2 Oropharyngeal sample collection/dimensions

Since 2014 the SHPVRL has offered an HPV genotyping service for the annotation of oropharyngeal cancers to all the Scottish NHS health boards. A total of 1798 samples with valid results were tested between 2013 and end of July 2020. Table 10 describes the number of samples received in each year of collection. Most oropharyngeal cancer(s) are squamous in origin. Therefore, morphology data was not presented for oropharyngeal samples.

**Table 10. Total number of oropharyngeal samples collected between 2013 and July 2020 – samples were received across 9 NHS health boards.**

Collection year	Total No cases	%
2013	158	8.79%
2014	230	12.79%
2015	261	14.52%
2016	234	13.01%
2017	229	12.74%
2018	262	14.57%
2019	302	16.80%
2020 (Up to July)	122	6.78%
<b>Total</b>	<b>1374</b>	-

### *2.5.6 Data verification/cleaning*

With respect to the data on cervical cancer for ambiguous database entries or missing information entries, contacts were made with PHS. Oropharyngeal data was obtained from the SHPVRL clinical database. Database “cleaning” & verification was performed to ensure quality of data. For database entries where it was not clear if the sample was oropharyngeal then the original request form was checked manually in addition to the laboratory electronic record. Ambiguous entries were checked with consultant staff and if the site was not specified the sample was excluded from the analysis. Most of the errors were related to spell variations in the anatomical site (i.e., oro, ORO, oropharyngeal) or classification of non-oropharyngeal samples as oropharyngeal (i.e., epiglottis or neck lymph node) Thus, filtering and rectification of erroneous details was completed.

### *2.5.7 Information capture for anal cancer and anal swab data (Chapters 4 and 6)*

#### *2.5.7.1 Anal cancer*

Anal cancer samples were not tested for HPV as part of routine diagnosis or surveillance program in Scotland. All anal lesions collected in the East of Scotland (Lothian, Borders and Fife regions) between 2009 and 2018 were requested. Biopsy samples were obtained as part of standard of care for the management of patients with anal disease; all biopsies had originally been obtained from the Southeast of Scotland NHS Lothian, Borders, and Fife) – these healthboards serve a population of 1,396,640 (Data from 2019)<sup>90</sup>.

A total of 224 anal samples, including high-grade and cancer lesions, were HPV genotyped. Anal cancer samples positive for HPV 16 were then tested for viral load, using a droplet digital PCR and HPV 16 sub-lineage identification through NGS.

### 2.5.7.2 Residual rectal swabs

To compare the sequences observed in the anal cancers, with a disease-free control group of anonymized residual rectal swabs obtained from asymptomatic men attending sexual health clinics were collated for downstream WGS. These samples had previously been genotyped as a consequence of immunization surveillance in Scotland. Only samples mono infected by HPV 16 were selected for NGS downstream (n=182).

## 2.6 Droplet digital PCR general steps

Viral load was analysed on invasive samples only. AIN (or high-grade) samples were not included in the analysis due to the small number of samples (n=15).

Invasive samples HPV 16 positive on the Seegene assay, were tested using a digital droplet PCR (ddPCR). Absolute quantification of viral load was performed on HPV 16-positive cancer samples (145 mono and 9 mixed infections) using a droplet digital assay (ddPCR). Nucleic acid was extracted from HPV 16-positive samples using the QIAamp DNA Mini Kit. (Qiagen, Germany) and sample concentration measurement was performed with the Qubit dsDNA High sensitivity kit (Thermo Fisher Scientific, MA, USA).

ddPCR was performed as described in Stevenson *et al.*, 2020, at the Centre for Virus Research, University of Glasgow. 0.7 µL of the RPP30 endogenous control assay, HPV 16 L1-specific primers and probes at 300 nM (final concentration) respectively, 10–100 ng of template DNA and 1 µL of restriction digest mix (consisting of 4 U of both EcoRI and HindIII in 1x NEB Cutsmart buffer (NEB, UK)) were used for the mix. Reactions were mixed with Droplet Generation Oil on DG8 cartridges in the QX200 droplet generator (Bio-Rad) to generate droplets. Thermal cycling conditions were: 95 °C for 10 min followed by 40×94°C for 30s and 60°C for 1min prior to final extension at 98°C for 10min.



Post-amplification, droplets were analysed on a QX200 Droplet Reader (Bio-Rad), and output data files were analysed using QuantaSoft analysis software v1.7.4 (Bio-Rad). The viral load for each sample was calculated relative to the endogenous RRP30 cellular gene internal control, with 2 copies present per cell. Any initially invalid results were repeated using a new FFPE section and fresh DNA extraction. After retesting, consistent invalids were not included in the analysis (n=9).

The individual HPV 16 viral loads were ranked from smallest to largest and separated using tertiles as described in Stevenson *et al.*, (2020). The VL threshold(s) for L1 low viral load was <12.3, medium between 12.3 to 57 and high viral load above 57.

## **2.7 General NGS methods used in chapter 6.**

HPV 16 sub-lineage identification using whole genome sequencing.

Due to the ratio HPV:human genome and potential fragmented DNA, the approach of sequencing chosen was target enrichment and short read sequencing.

### *2.7.1 Target enrichment*

HPV 16 DNA was amplified using a conventional PCR. The entire HPV 16 genome (7906 bp) was amplified by 47 overlapping amplicons, ranging in size from 181 bp to 375 bp, as described previously (Cullen, *et al.* 2015<sup>32</sup>) and optimised by Arroyo *et al.* 2018<sup>52</sup>. Primer sets were divided into five different reactions to decrease self-dimer and cross-primer dimer formation (Table 11).

**Table 11. Primers pools for HPV 16 sequencing.**

POOL1	POOL2	POOL3	POOL4	POOL5
HPV 16_7_F	HPV 16_1_F	HPV 16_2_F	HPV 16_19_F	HPV 16_10_F
HPV 16_7_R	HPV 16_1_R	HPV 16_2_R	HPV 16_19_R	HPV 16_10_R
HPV 16_9_F	HPV 16_3_F	HPV 16_4_F	HPV 16_22_F	HPV 16_12_F
HPV 16_9_R	HPV 16_3_R	HPV 16_4_R	HPV 16_22_R	HPV 16_12_R
HPV 16_13_F	HPV 16_5_F	HPV 16_6_F	HPV 16_24_F	HPV 16_21_F
HPV 16_13_R	HPV 16_5_R	HPV 16_6_R	HPV 16_24_R	HPV 16_21_R
HPV 16_15_F	HPV 16_14_F	HPV 16_8_F	HPV 16_28_F	HPV 16_23_F
HPV 16_15_R	HPV 16_14_R	HPV 16_8_R	HPV 16_28_R	HPV 16_23_R
HPV 16_15B_R	HPV 16_14B_R	HPV 16_8B_F	HPV 16_28B_R	HPV 16_26_F
HPV 16_17_F	HPV 16_16_F	HPV 16_8B_R	HPV 16_32_F	HPV 16_26_R
HPV 16_17_R	HPV 16_16_R	HPV 16_11_F	HPV 16_32_R	HPV 16_30_F
HPV 16_20_F	HPV 16_31_F	HPV 16_11_R	HPV 16_34_F	HPV 16_30_R
HPV 16_20_R	HPV 16_31_R	HPV 16_18_F	HPV 16_34_R	HPV 16_33_F
HPV 16_20B_F	HPV 16_31B_R	HPV 16_18_R	HPV 16_43_F	HPV 16_33_R
HPV 16_25_F	HPV 16_36_F	HPV 16_29_F	HPV 16_43_R	HPV 16_39_F
HPV 16_25_R	HPV 16_36_R	HPV 16_29_R	HPV 16_45_F	HPV 16_39_R
PV16_27_F	HPV 16_38_F	HPV 16_29B_F	HPV 16_45_R	
HPV 16_27_R	HPV 16_38_R	HPV 16_37_F		
HPV 16_35_F	HPV 16_42_F	HPV 16_37_R		
HPV 16_35_R	HPV 16_42_R	HPV 16_40_F		
HPV 16_41_F	HPV 16_46_F	HPV 16_40_R		
HPV 16_41_R	HPV 16_46_R			
HPV 16_44_F				
HPV 16_44_R				
HPV 16_47_F				
HPV 16_47_R				

*2.7.2 Concentration primers:*

Primers were ordered as 25 nmole DNA. When received all 108 primers, were diluted to 100µM using molecular DNA/RNA free water (Sigma Aldrich, USA) and stored a -20°C. Pools with primers described in table 11 where prepared and diluted as follow:

- Pool 1: 260ml of primers

- Pool 2: 220 ml of primers + 40 ml of DNA/RNA free water
- Pool 3: 210 ml of primers + 50 ml of DNA/RNA free water
- Pool 3: 170 ml of primers + 90 ml of DNA/RNA free water
- Pool 3: 160 ml of primers + 100 ml of DNA/RNA free water

### *2.7.3 PCR mastermix and amplification*

PCRs were performed using Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany) and 0.2  $\mu$ M of each primer, according to manufacturers' instructions:

- 12.5  $\mu$ l of Qiagen Multiplex PCR
- 2.5  $\mu$ l of 10  $\mu$ M primers pool
- 5  $\mu$ l of molecular DNA/RNA free H<sub>2</sub>O.
- 5  $\mu$ l of Sample
- **Total volume of 25  $\mu$ l**

PCR was completed with a total volume of 25  $\mu$ l, using an Applied Biosystem 9700 thermocycler (MA, USA). Cycling conditions selected were pre-heat for 15 min at 95 °C, 45 cycles at 95°C for 30 seconds, 57°C for 90 seconds and 72°C for 90 seconds, with a final extension at 72 °C for 10 min. PCR amplification products were then pooled together according to sample name prior to library preparation and stored at 4°C for next-day library preparation.

### *2.7.4 Library Preparation*

#### *2.7.4.1 Illumina TruSeq DNA nano library*

Libraries were originally prepared using the protocol used at the HPV International Laboratory (Karolinska Institute, Sweden), which required the use of the Illumina TruSeq DNA nano library kit. However, this protocol was optimised by Arroyo *et al*, 2018<sup>52</sup> for small size amplicons, starting from the A-tail step and avoiding the fragmentation and end-repair steps.

The TruSeq kit requires to perform a normalisation step. This step consists in diluting the different libraries to the same concentration before pooling them to ensure an even distribution of DNA for all the samples.

#### *2.7.4.2 Illumina DNA Prep Kit*

Libraries for anal cancer and control cohort samples were then prepared using the Illumina DNA prep kit (San Diego, CA, USA) following the manufacturer's instructions, using 450 ng of DNA in 35 µL as input. Sequencing was performed using the Illumina MiSeq instrument and the Illumina MiSeq reagent kit v2 500 cycles (2 × 250 bp). Libraries were normalized to 4 nM in combination with 12.5 pM of PhiX (Illumina).

A total volume of 594 µl of denatured library (20pM) + 6 µl of PhiX (12.5pM) was added to the Illumina MiSeq reagent kit v2 cartridge. MiSeq run was configured as FASTQ only. Sample sheet was created using the Illumina Experiment manager, selecting number of cycles and indexes.

#### *2.7.5 Bioinformatic Analysis*

Reads obtained from Illumina were de-multiplexed and converted to fastq files. All fastq files were quality and adaptor trimmed using Trimmomatic (v0.39)<sup>91</sup>. Only high-quality paired reads (-phred 33 -leading 3 -trailing 3- slidingWindow: 4:15) with 150 bp were used for further analysis. FASTQC tools were further used to assess whether any adaptors remained<sup>92</sup>. High-quality reads were then mapped to the HPV 16 reference genome from the Papillomavirus Episteme<sup>93</sup> using bwa (v0.7.17)<sup>94</sup>, to create a sam file. Due to

the circular HPV genome, the reference genome was modified by adding the 258 nucleotides from the beginning to the end of the genome sequence to not lose coverage of amplicons 46 and 47. SAMtools (v1.14)<sup>95</sup> was then used to convert files from sam to bam and to curate files for the variant calling. BCFtools (v1.14), mpileup and consensus tools were used for the variant calling and for the generation of a consensus sequence<sup>96</sup>, using default parameters. Positions not covered were annotated as Ns.

New consensus files were aligned using MAFFT (v7.490)<sup>97</sup> with default parameters. A manual edit was performed when required. Maximum likelihood trees were inferred using RAxML (v2.0.8)<sup>97,98</sup> with the GTR substitution model (ML + transfer bootstrap expectation + consensus, 1 run, 100 reps). Visualization of the trees generated by RAxML was performed using Figtree (v1.4.4). Each sample was assigned with a sub-lineage corresponding to the nearest neighbour.

Sub-lineages references were obtained from the PAVE for each of the HPV 16 sub-lineages: A1 (Accession number [K02718.1](#)), A2 (Accession number [AF536179.1](#)), A3 (Accession number [HQ644236.1](#)), A4 (Accession number [AF534061.1](#)), B1 (Accession number [AF536180.1](#)), B2 (Accession number [HQ644298.1](#)), B3 (Accession number [HQ644298.1](#)), B4 (Accession number [KU053914.1](#)), C1 (Accession number [AF472509.1](#)), C2 (Accession number [HQ644244.1](#)), C3 (Accession number [KU053920.1](#)), C4 (Accession number [KU053925.1](#)), D1 (Accession number [HQ644257.1](#)), D2 (Accession number [AY686579.1](#)), D3 (Accession number [AF402678.1](#)) and D4 (Accession number [AF402678.1](#)) A sub-lineage assignment was performed for all specimens excluding those with <100× median depth or low genome coverage (<80% genome coverage).

### *2.7.6 Validation of the library preparation and bioinformatic analysis*

As a further quality control for analysis of the sequence data generated, a subset of anonymised 25 fastq files were sent to the International HPV Reference Laboratory in Karolinska, Sweden for independent and blind bioinformatic analysis and sub-lineage identification. The Karolinska pipeline included the following tools: BCL2FASTQ, Trimmomatic, NextGenMap, GATK, Trimbam, Fixmate and MEGA.

### *2.7.7 Assessment of HPV integration in anal cancer samples*

Coverage of all samples were analysed using Artemis<sup>99</sup> and Qualimap<sup>100</sup>. Samples with regions with no sequencing reads were repeated from the PCR step to discard any potential error during the process.

Annotation of the plausible integration regions was performed using Artemis<sup>99</sup> and annotation file from HPV 16 A1 reference (K02718.1).

## **2.8 Statistical packages**

Statistical analysis was performed using R Studio (2022.07.2, build 576) with the following packages:

- Survival: Survival Analysis (version 3.5.3)<sup>101</sup>
- Survminer: (version 0.4.9)<sup>101</sup>
- ggplot2 (version 3.4.1)<sup>102</sup>
- questionr (version 0.7.4)

## 2.9 Primers

The entire HPV 16 genome (7906 bp) was amplified as 47 overlapping amplicons, ranging in size from 181 bp to 375 bp, as described previously (Cullen, *et al.* 2015). Primers were ordered from Eurogentec ([www.eurogentec.com](http://www.eurogentec.com))

First set of primers (12 primers)

>HPV 16_7_F	GCTGCAAAAAGGAGATTATTTG
>HPV 16_7_R	ATTGCTGCCTTTGCATTACT
>HPV 16_9_F	GGTGTATTGCTGCATTTGGA
>HPV 16_9_R	TACGCAATTTTGGAGGCTCT
>HPV 16_13_F	AGATGTGATAGGGTAGATGATGGAG
>HPV 16_13_R	TTGTCATCTATGTAGTTCCAACAGG
>HPV 16_15_F	GATTGGTGGTGTTTACATTTCC
>HPV 16_15_R	CATTCTAGGCGCATGTGTTT
>HPV 16_15B_R	CATTCTAGGCGCATTTGTTT
>HPV 16_17_F	GTGCCAACACTGGCTGTATC
>HPV 16_17_R	TGCATATGTCTCCATCAAAGTG
>HPV 16_20_F	TCTGTGTTTAGCAGCGACGA
>HPV 16_20_R	CAGTGAGGATTGGAGCACTG
>HPV 16_20B_F	TCTGTGTTTAGCAGCAACGA
>HPV 16_25_F	GGATAACAGCGGCCTCTGC
>HPV 16_25_R	AAAGTTGGGTAGCCGATGC
>PV16_27_F	GTACAGGCGGACGCACTG
>HPV 16_27_R	GGGATTATTATGTGTAGTAACAGTAGTAACAG
>HPV 16_35_F	CTTGCAGTTGGACATCCCTA
>HPV 16_35_R	CACACCTAATGGCTGACCAC
>HPV 16_41_F	TGTGCAAAATAACCTTAACTGC
>HPV 16_41_R	TGCGTCCTAAAGGAAACTGA
>HPV 16_44_F	GTTTGTATGTGCTTGTATGTGCTTG
>HPV 16_44_R	CGGTTGAAGCTACAAAATGGC

>HPV 16\_47\_F           CAAACCGTTTTGGGTTACAC  
>HPV 16\_47\_R           ATGCATAAATCCCGAAAAGC

Self-Dimers:           1 dimer for: HPV 16\_25\_F  
                          1 dimer for: HPV 16\_47\_F  
                          1 dimer for: HPV 16\_44\_R

Cross Primer Dimers:  HPV 16\_25\_F with HPV 16\_25\_R  
                          HPV 16\_9\_F with HPV 16\_44\_R  
                          HPV 16\_9\_R with HPV 16\_44\_R  
                          HPV 16\_13\_R with HPV 16\_44\_R  
                          HPV 16\_20\_F with HPV 16\_25\_F

Second set of primers (10 primers)

>HPV 16\_1\_F           ACAGTTACTGCGACGTGAGG  
>HPV 16\_1\_R           TGGAATCTTTGCTTTTTGTCC  
>HPV 16\_3\_F           TTGCAACCAGAGACAACTGA  
>HPV 16\_3\_R           TTCTGAGAACAGATGGGGCAC  
>HPV 16\_5\_F           ACGGGATGTAATGGATGGTT  
>HPV 16\_5\_R           TGTTGTTTTGCTTCCTGTGC  
>HPV 16\_14\_F          GCAGATGCCAAAATAGGTATG  
>HPV 16\_14\_R          ACTGGATTTCCGTTTTTGTC  
>HPV 16\_14B\_R         ACTGGATTTCCGTTTTCGTC  
>HPV 16\_16\_F          AAAACGATGGAGACTCTTTGC  
>HPV 16\_16\_R          AGTTGCAGTTCAATTGCTTGT  
>HPV 16\_31\_F          GCTCCAGATCCTGACTTTTTG  
>HPV 16\_31\_R          AGTIGGTGAGGCTGCATGGG  
>HPV 16\_31B\_R         AGTIGGTGAGGCTGCATGTG  
>HPV 16\_36\_F          GTTTGGGCCTGTGTAGGTG  
>HPV 16\_36\_R          TTCCCCTATAGGTGGTTTGC  
>HPV 16\_38\_F          CGGCTTTGGTGCTATGGAC  
>HPV 16\_38\_R          GCAGTAGACCCAGAGTCTTTAATG



>HPV 16\_42\_F           GGAGGCACACTAGAAGATACTTATAGG  
>HPV 16\_42\_R           GAGGTGGTGGGTGTAGCTTTTC  
>HPV 16\_46\_F           TAAATTACTATGCGCCAACG  
>HPV 16\_46\_R           ACTAACCGGTTTCGGTTCAA

Self-Dimers:           1 dimer for: HPV 16\_36\_R  
                          1 dimer for: HPV 16\_42\_F  
                          2 dimers for: HPV 16\_38\_R

Cross Primer Dimers: HPV 16\_14\_F with HPV 16\_14\_R  
                          HPV 16\_36\_F with HPV 16\_36\_R  
                          HPV 16\_14\_F with HPV 16\_42\_F  
                          HPV 16\_31\_F with HPV 16\_38\_F

Third set of primers (9 primers)

>HPV 16\_2\_F           TCAAAAGCCACTGTGTCCTG  
>HPV 16\_2\_R           TTCATCCTCCTCCTCTGAGC  
>HPV 16\_4\_F           GCGTACAAAGCACACACGTA  
>HPV 16\_4\_R           CTGTCATTTTCGTTCTCGTCA  
>HPV 16\_6\_F           TTTAACACAGGCAGAAACAGAGAC  
>HPV 16\_6\_R           CGCCCTTCTACCTGTAACATC  
>HPV 16\_8\_F           TGCGAAACACCACTTACAAA  
>HPV 16\_8\_R           CATTCCCCATGAACATGCTA  
>HPV 16\_8B\_F          TGCCAAACACCACTTACAAA  
>HPV 16\_8B\_R          CATTCCCCATGAACACGCTA  
>HPV 16\_11\_F          ACACGCCAGAATGGATACAA  
>HPV 16\_11\_R          CACATTGTTGCACAATCCTTT  
>HPV 16\_18\_F          AAACATGGATATACAGTGAAGTGC  
>HPV 16\_18\_R          ATTACCTGACCACCCGCATG  
>HPV 16\_29\_F          AGCACAAACCCTAACACAGTAACTAG  
>HPV 16\_29\_R          TAAAATGTATTATCCACATCTATACCTTC

>HPV 16\_29B\_F AGCACAAATCCTAACACAGTAACTAG  
>HPV 16\_37\_F AGCAAATGCAGGTGTGGATA  
>HPV 16\_37\_R TCCAGTGGAACCTTCACTTTTG  
>HPV 16\_40\_F ATGGCATTGTGGGGTAAC  
>HPV 16\_40\_R ACCAAAATTCAGTCCTCCA

Self-Dimers: 1 dimer for: HPV 16\_8\_R  
1 dimer for: HPV 16\_29\_R

Cross Primer Dimers: HPV 16\_6\_F with HPV 16\_6\_R  
HPV 16\_40\_F with HPV 16\_29\_F  
HPV 16\_37\_F with HPV 16\_29\_R

#### Fourth set of primers (8 primers)

>HPV 16\_19\_F TGCAGTTTAAAGATGATGCAGA  
>HPV 16\_19\_R CGCTGGATAGTCGTCTGTGT  
>HPV 16\_22\_F CCATAGTACATTTAAAAGGTGATGC  
>HPV 16\_22\_R CGCCAGTAATGTTGTGGATG  
>HPV 16\_24\_F GTGTGCTTTTGTGTGTCTGCC  
>HPV 16\_24\_R TGTGTCGCATTGTTAAGTGATAAC  
>HPV 16\_28\_F TCAACTGATACCACACCTGCT  
>HPV 16\_28\_R GAGACCCTGGTATGGGTGTG  
>HPV 16\_28B\_R GAGACCCCGGTATGGGTGTG  
>HPV 16\_32\_F GTAGAATTGGTAATAAACAAACTACG  
>HPV 16\_32\_R GGAAGTAATGAAGGAGTTTGGTCAG  
>HPV 16\_34\_F CATGTTACGAAAACGACGTAAC  
>HPV 16\_34\_R CCAAACCTTATTGGGGTCAGG  
>HPV 16\_43\_F AAGGCCAAACCAAATTTACA  
>HPV 16\_43\_R GCATGACACAATAGTTACACAAGC  
>HPV 16\_45\_F GCCATTTTGTAGCTTCAACCG  
>HPV 16\_45\_R CAAGCCAAAAATATGTGCCTAAC

Self-Dimers: 1 dimer for: HPV 16\_45\_F  
Cross Primer Dimers: HPV 16\_22\_F with HPV 16\_43\_F  
HPV 16\_28\_F with HPV 16\_28\_R  
HPV 16\_28\_F with HPV 16\_28B\_R  
HPV 16\_32\_F with HPV 16\_32\_R

Fifth set of primers (8 primers)

>HPV 16\_10\_F TGTGTGTCTCCAATGTGTATGATG  
>HPV 16\_10\_R CCCATTGTACCATCGTGATAATTC  
>HPV 16\_12\_F TGCACAATTGGCAGACACTA  
>HPV 16\_12\_R ACCTGTGTTAGCTGCACCAT  
>HPV 16\_21\_F ACCCCTGCCACACCAATAAG  
>HPV 16\_21\_R TATGTCCTGTCCAATGCCATG  
>HPV 16\_23\_F TACACTTACATATGATAGTGAATGTCAACG  
>HPV 16\_23\_R CGTATGTAGACACAGACAAAAGCAG  
>HPV 16\_26\_F GTTCTGCAAAACGCACAAA  
>HPV 16\_26\_R GGGGTCTTACAGGAGCAAGT  
>HPV 16\_30\_F GACCCTGCTTTTGTAACTC  
>HPV 16\_30\_R GTACGCCTAGAGGTTAATGCTGG  
>HPV 16\_33\_F TTCCTGCAAATACAACAATTCC  
>HPV 16\_33\_R TACTGGGATAGGAGGCAAGTAGAC  
>HPV 16\_39\_F TGGTGAAAATGTACCAGACGA  
>HPV 16\_39\_R GATATGGCAGCACATAATGACA

Self-Dimers: 1 dimer for: HPV 16\_10\_R  
1 dimer for: HPV 16\_26\_F

Cross Primer Dimers: HPV 16\_30\_F with HPV 16\_30\_R  
HPV 16\_21\_R with HPV 16\_39\_R  
HPV 16\_10\_F with HPV 16\_23\_R  
HPV 16\_23\_R with HPV 16\_30\_F

### 3. HPV type specific prevalence in cervical and oropharyngeal cancer in the Scottish population.

#### 3.1 Introduction

As described in the introduction, high-risk HPV types (hr-HPV) are causative of carcinogenic lesions in different anatomical locations, including cervix, oropharynx, anus, penis, vagina, and vulva. The most common HPV-driven cancers in Scotland are cervical and oropharyngeal cancer. Anal cancer being the third most common, (ISD, data extracted May 2019)<sup>84</sup>. Furthermore, evidence suggests that oropharyngeal and anal cancer is rising in Scotland. Figure 8 shows the EASR incidence from 1993 to 2017. While cervical cancer incidence per 100,000 has decreased from 2015, other HPV-driven cancers like oropharyngeal, anal, or penile cancer have increased since 1993.

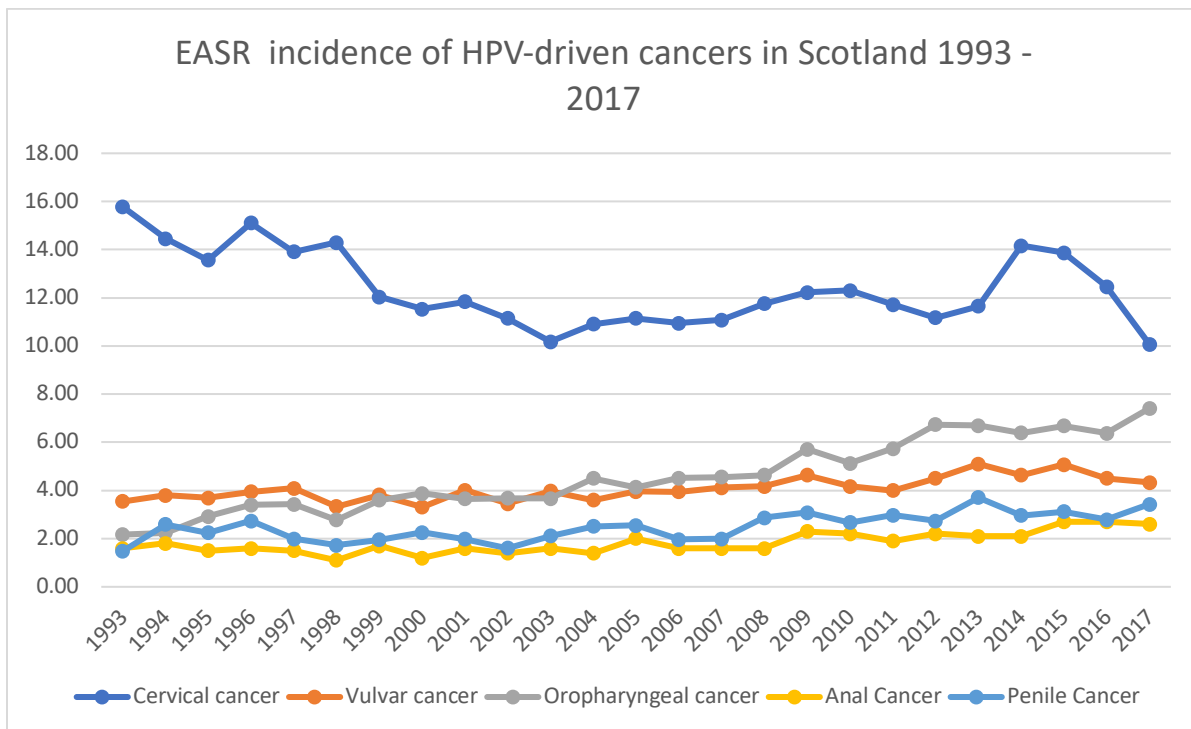


Figure 8. EASR: age-standardised incidence rate per 100,000 person-years at risk of the 5 most common HPV-driven cancers in Scotland.

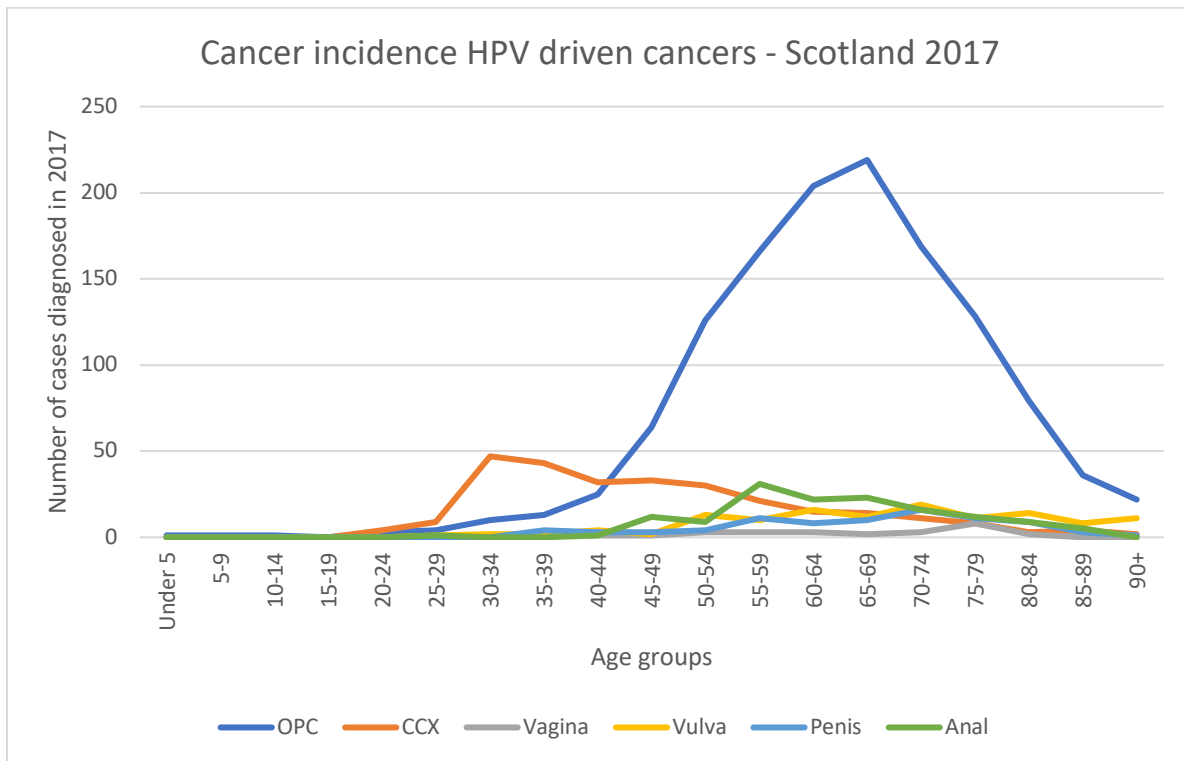
Different studies have been performed in Scotland in the past looking at HPV prevalence in cervical and oropharyngeal lesions. In 2010, Cuschieri *et al.* analysed the prevalence of HPV in cervical cancer (CCX) biopsies collected between 2004 to 2008<sup>103</sup>. Samples were collected in the main pathology centres in Scotland. A high prevalence of HPV high risk types 16 and 18 were present in 82% of the HPV positive cervical biopsies analysed. The next most frequently detected types in the Scottish invasive cervical cancer (ICC) cases were, in order of prevalence, HPV 45, 33 and 31. However, these data are now 10 years old, and a more contemporary epidemiological data set could help provide a more accurate depiction of clinically relevant types in Scotland and whether these have changed.

As Figure 8 shows, oropharyngeal squamous cell carcinoma (OPSCC) is one of the cancers which has increased dramatically in recent years, not only in the UK, but also in the developed countries, existing a big difference in incidence with LMIC<sup>104</sup>.

The United Kingdom has seen a 51% increase in oral and oropharyngeal squamous cell carcinoma in men from seven per 100000 to eleven per 100000 between 1989 and 2006<sup>8</sup>. While not all the oropharyngeal cancers are driven by HPV - a UK study of cancers collected from 2002 to 2011, indicated that approximately 50-55% of the cases were HPV positive<sup>105,106</sup>. In Scotland, two studies looked at the HPV positivity rate in oropharyngeal cancer. Wells *et al.*, 2015 looked at oropharyngeal cancer (OPC) in the Southeast of Scotland. The authors found HPV in 57% of OPCs, with HPV 16 type dominating in 90% of the HPV positive cases. Wakeham *et al.*, 2019 looked at the HPV-driven OPC in the West of Scotland, with an HPV prevalence of 60% on a cohort of cases diagnosed between 2013-15<sup>107</sup>. HPV 16 was found in 55.5% of samples. This latter publication did not consider the influence of age on prevalence and the sample was “restricted” to the West of Scotland.

Nonetheless, most of the Scottish and UK data is now over 10 years old now, and a newer data set could help providing a more contemporary assessment of the HPV types in the OPC population and help model the impact of primary and secondary interventions including vaccination.

Incidence of HPV-driven cancers varies by age group (Figure 9). It would be interesting to see if HPV attributable fraction of cervical and oropharyngeal cancers vary depending on age<sup>84</sup>.



**Figure 9. Incidence of HPV-driven cancers detected in Scotland in 2017.** Data obtained from Information Division Service, Scotland. Accessed 02/04/23.

One of the key aims of the first part of this thesis is to identify the extent and diversity of HPV types in Scotland associated with the most common HPV driven cancers (cervical and oropharynx). Additionally, as mentioned in the introduction, there are rare cervical cancers, where HPV is not detected. However, thanks to the use of next generation sequencing, it has been identified that a large number of cases classified as HPV negative, were indeed positive for HPV<sup>61,62</sup>. By analysing the HPV status on the cervical

cancers described in this chapter, I will be able not only to determine what is the vaccine preventable component of cervical and oropharyngeal cancers, but also determining the extent and nature of the remaining portion. By understanding better the HPV negative component of the cancers, we could try and tailor prevention and management strategies in the future.

To accomplish this, an interrogation into type specific prevalence, with particular respect to high-risk HPV types (HR-HPV) (including non-vaccine and cross-reactive types) were performed for the following:

- Cervical cancer diagnosed in 2015, 2016 and 2017.
- Oropharynx cancer diagnosed between 2013 and 2018.

### **3.2 Overarching Aim:**

This chapter aims to provide a contemporary description of HPV types in the most common invasive lesions in Scotland to define what types are clinically relevant for further study/interrogation now, and in the future given the impact of vaccination. The specific aims to this chapter are:

- To assess the type specific diversity of HPV types in cervical cancer and oropharyngeal cancer in Scotland and identify any changes over time.
- To ascertain if there are differences in the most frequently occurring types according to geography/health board of diagnosis in cervical cancers.
- To investigate the association of HPV status (presence or absence) with demographic variables (age, health board, etc).
- To describe the preventable fraction of the 2 cancer lesions driven by HPV by the current licensed HPV vaccines.



### 3.3 Material and Methods

#### 3.3.1 Collection period for cervical cancer samples

A total of 649 cervical invasive lesion samples were sent to the Scottish HPV Reference Laboratory (SHPVRL) for HPV genotyping from the 3-collection years (2015, 2016 and 2017), described in Table 12. Histological findings were classified according to the British Association for Cytopathology and NHS cervical screening program criteria<sup>87-89</sup>. Cervical cancer and samples were classified in 3 different groups attending to their histology identification: Squamous cell carcinoma (SCC), adeno squamous carcinoma (ASC) and adenocarcinoma (ADC).

**Table 12. Total number of high-grade and cervical invasive lesions collected, stratified by morphology.**

Samples were collected as part of a national HPV surveillance exercise from the Scottish NHS health boards.

	Cervical cancer			
Diagnose year	SCC	ASC	ADC	Unknown
2011	N/A	N/A	N/A	N/A
2013	N/A	N/A	N/A	N/A
2015	176	13	50	5
2016	112	1	32	77
2017	68	0	26	81
<b>Total</b>	356	14	108	163

Eight samples could not be tested for HPV as the genotyping result was invalid and therefore not included in the analysis.

### *3.3.2 Morphology and histology of cervical cancer*

Table 1 represents the number of cases stratified by histology and morphology. 55.16% (n=356) of samples received were squamous cervical carcinoma (SCC), 2.16% (n=14) adenosquamous carcinoma (ASC), 16.64% (n=108) adenocarcinoma (ADC), while 25.11% (n=163) unknown. ASC and ADC were aggregated in one group for the regression analysis due to the small number of ASC cases.

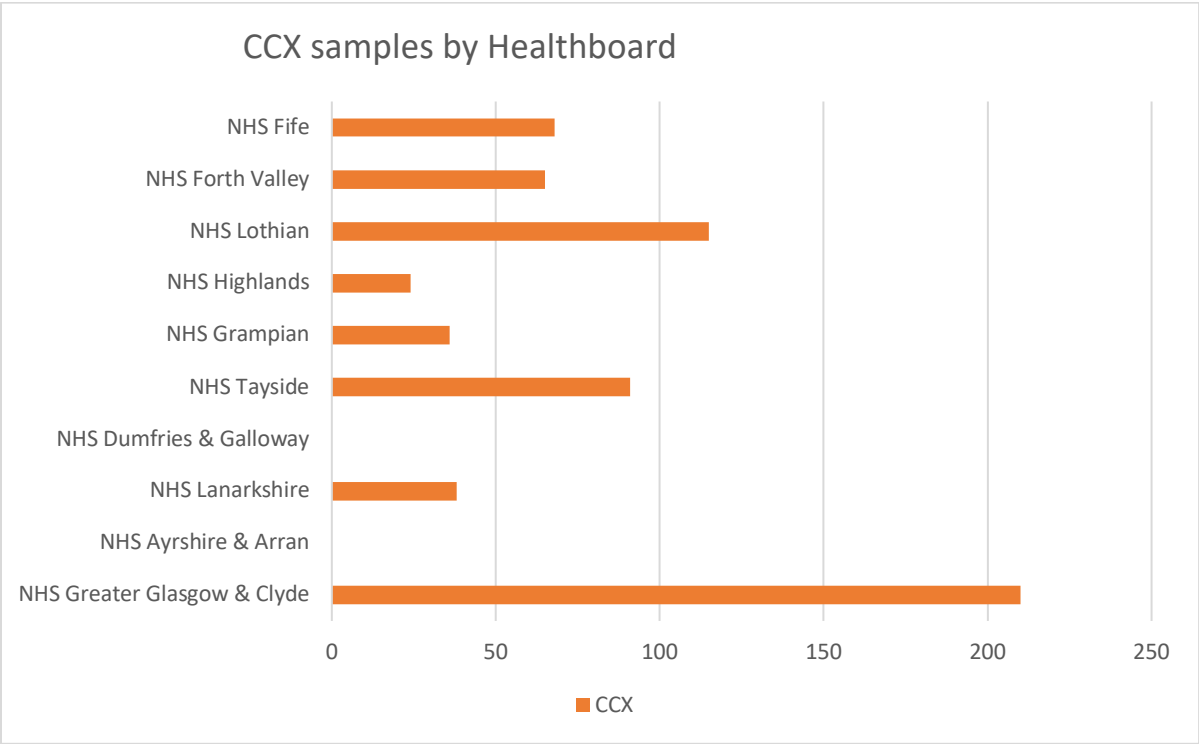
### *3.3.3 Demographics*

#### *3.3.3.1 Age- Cervical cancer*

Full data are presented in appendix 1. Some age information was not available for all the samples. 154/641 (24.02%) of samples did not contain age information. When it was available, age of patients was separated into 5 groups: <45, 45 – 54, 55 – 64, 65 – 74 and >75. The age range was between 22 to 95 years old. The greatest number of samples collected was from age group <45 with 36.66% of samples followed by those in the 45 - 55 age group (18.25%).

#### *3.3.3.2 Health board – Cervical cancer*

Cases stratified by healthboards is represented in Figure 10. Cervical samples were requested from health boards, according to catchment area. Consequently, the largest number of samples were collected by Greater Glasgow and Clyde health board 210/641 (32.76%) – the largest territorial health board in Scotland - followed by Lothian 115/641 (17.94%) and Tayside 91/641 (14.20%), 68/641 Fife (10.61%), 65/641 Forth Valley (10.14%), 38/641 Lanarkshire (5.93%), 36/641 Grampian (5.62%) and Highlands 24/641 (3.74%).



**Figure 10. Number of cervical cancer samples collected in each Scottish NHS healthboards.**

*3.3.3.3 Collection years – Oropharyngeal samples*

A total of 1798 samples with valid results were tested between 2013 and end of July 2020. Table 13 describes the number of samples received in each year of collection. A valid result included the presence or absence of HPV and the detection of the internal control.

**Table 13. Total number of oropharyngeal samples collected between 2013 and July 2020 – samples were received across 9 health boards.**

Collection year	Total No cases	%
2013	158	8.79%
2014	230	12.79%
2015	261	14.52%
2016	234	13.01%
2017	229	12.74%
2018	262	14.57%
2019	302	16.80%
2020 (Up to July)	122	6.78%
<b>Total</b>	<b>1798</b>	-

Most oropharyngeal cancer(s) are squamous in origin, and it is the squamous cancers that are generally sent for testing to SHPVRL. Morphology data is not presented for oropharyngeal samples.

### 3.3.4 Demographic – Oropharyngeal samples

#### 3.3.4.1 Age and sex

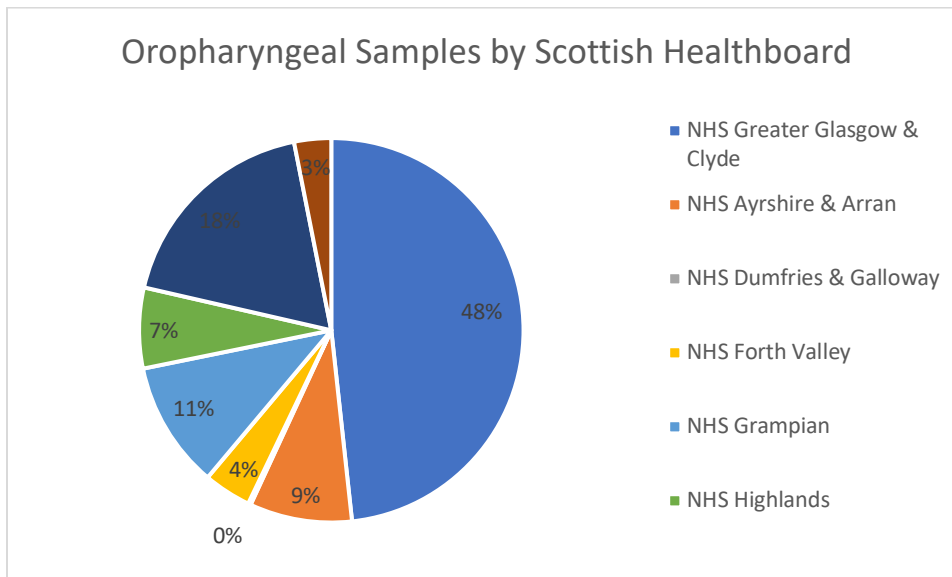
The median age of patients to oropharyngeal cases was 61.38 (SD ± 11.13). Majority of cases were in the 60 – 69 years old group (32.98%) followed by the 50 - 59 group with 32.81%. The <50 accounted for 12.06% of samples and the ≥70 for 21.91%.

#### 3.3.4.2 Sex

A total of 1798 samples collected were from men (73.36%) and 479 from female (26.64%).

### 3.3.4.3 Health board – Sample location

Less healthboards/laboratories are represented for the OPC collection compared to those represented in the cervical surveillance work given that surveillance was a national prospective exercise whereas HPV testing of OPC relates to a service which although offered nationally receives samples from “centres” that collect samples and perform pathology diagnosis from more than one board. Also, the service may be accessed differentially depending on local protocols. Figure 11 shows the percentage of samples received from health boards between 20013 to 2020. NHS Greater Glasgow & Clyde, the largest of all Scottish Health boards and is the healthboard which sent most samples (48%)



**Figure 11. NHS Health board distribution of oropharyngeal samples received between 20013 – 2020.**

Frequency of HPV infections were counted and ranked in order of prevalence – an infection was counted if it occurred as a mono infection or as a multiple infection with other types.

### *3.3.4 Statistical analysis*

#### *3.3.4.1 Statistical analysis – cervical cancer*

To assess the relationship in cervical cancers between HPV positivity and different factors (two or more independent variables) a univariate logistic regression analysis approach was performed between overall HPV result (any HPV positive) and age at diagnosis, collection year, morphology, and health board of diagnosis. Adjustment was performed for age, morphology, and health board. Moreover, due to the dominance of types 16 and 18, the same analysis was also performed focusing on these types as a duo. Odds ratio (OR) were calculated to quantify the strength of the association between HPV positivity and the different demographic and clinical data. All the statistics were obtained using R-studio macOS, (version 1.2.1335).

#### *3.3.4.2 Statistical analysis – Oropharyngeal cancer*

For the oropharyngeal cancer, the analysis approach was very similar to above. To assess the relationship in oropharyngeal cancers between HPV positivity and different factors (two or more independent variables) a univariate logistic regression analysis approach was performed between overall HPV result (any HPV positive) and age at diagnosis, collection year, sex, and health board of diagnosis. Adjustment was performed for age, sex, and health board. Odds ratios (OR) were calculated to calculate the strength of the association between HPV positivity and the different demographic and clinical data. All the statistics were obtained using R-studio macOS, (version 1.2.1335).

## 3.4 Results

### 3.4.1 HPV type prevalence in cervical cancer

#### 3.4.1.1 Overall HPV type prevalence in invasive cervical samples

A total of 649 FFPE cervical cancer samples were tested for HPV in the present study; 8 had an invalid HPV result and were not included in the analysis (n=641). Two hundred and forty-four (38.1%) samples were collected in 2015, 222 (34.63%) in 2016 and 175 (27.30%) in 2017.

#### 3.4.1.2 Overall HPV type positivity

Type specific prevalence is shown in detail in Appendix 1. Of the 641 cases, 587 samples tested positive (91.58%, 95% CI 89.17 – 93.49) for at least one of the 24 HPV types detected by the genotyping assay. A total of 8.42% (95% CI 6.51- 10.83) of samples were HPV-negatives. High-risk types were detected in 577 samples (90.02%, 95% CI 87.46 – 92.11). HPV 16 and/or 18 was detected in 490 samples (75.50%, 95% CI 72.05 – 78.65)

Type HPV 16 was the most prevalent with 394 samples positive (60.71%, 95% CI 56.90 – 64.39). HPV 18 DNA was the second most common, detected in 117 samples (18.03%, 95% CI 15.26 – 21.17). Other hr-types were positive in 35 samples (5.46%, 95% CI 3.95 – 7.50) and non-hr-types (including low risk types) were positive in only 10 samples (1.54%, 95% CI 0.84 – 2.81). When considering the 5 most prevalent types, the order was HPV 16>HPV 18>HPV45>HPV31>HPV33.

As a comparison with a previous study, Cuschieri et al., 2010<sup>68</sup> that reflected Scottish cervical cancer cases collected before 2004, where overall positivity was very similar (88%) to the overall HPV positivity identified within this study (91.58%). In the same way, 16 and/or 18 positivity from this study and the

2010 study reported similar positivity, with 75.50% (95% CI 72.05 – 78.65) and 72% (95% CI 67 – 76) respectively in the overall cases.

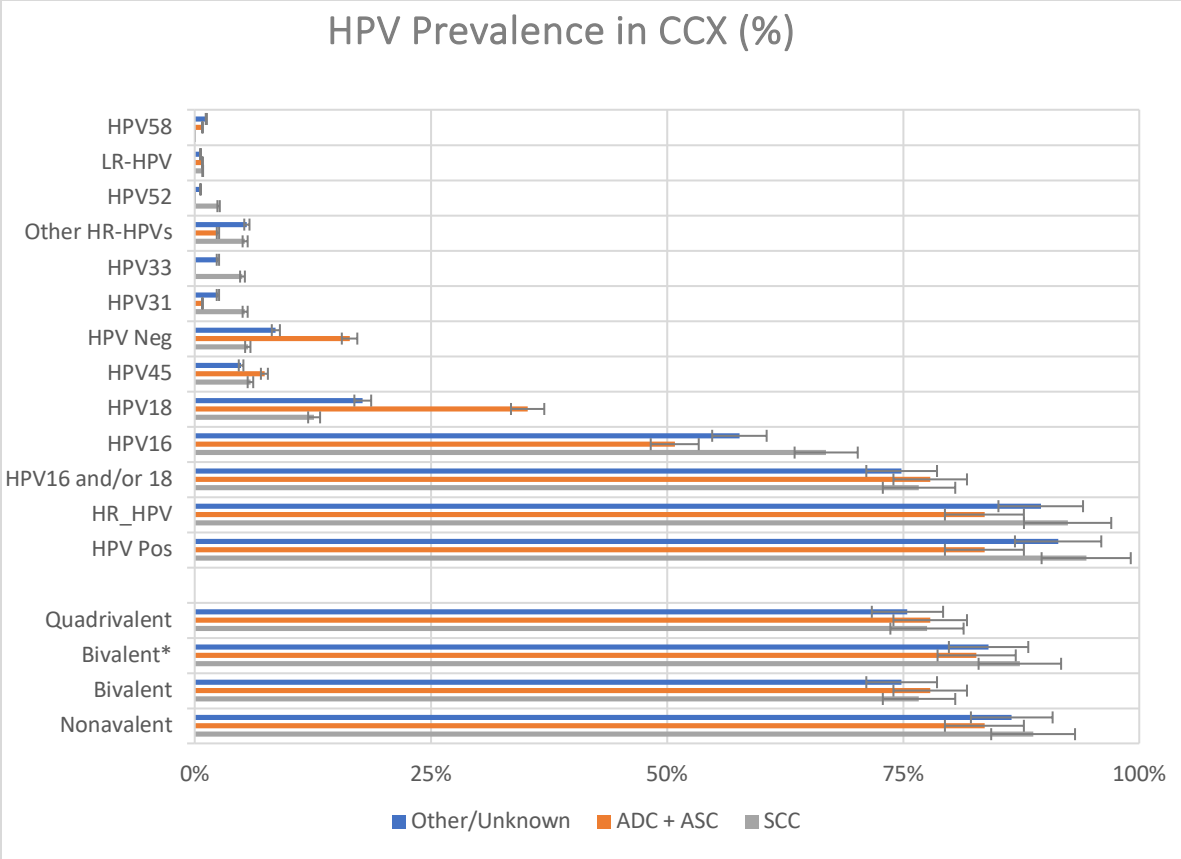
#### *3.4.1.3 HPV type positivity in cervical cancer according to morphology*

As mentioned in the methodology, cervical cancers differ according to underlying morphology. Morphology information was available for 478/641 (74.6%) while no information was available for 163 (25.4%). In addition, due to the small number for ADC and ADC they were aggregated for higher statistical weight. Figure 12 describes the different HPV prevalence by morphology.

A total of 336 samples (94.38%, 91.48 - 96.33) of SCC samples were positive for any HPV type, while 102/122 (83.61%, 76.04 - 89.13) of ASC+ADC were HPV positive. When focussing on “just” hr-HPV types, samples were positive in 91.90% (CI 88.61 – 94.30) and 83.61% (CI 76.04 – 89.13) of samples for SCC, ASC + ADC respectively. When overall HPV and HR-HPV prevalence of ASC + ADC is compared against SCC, ASC+ADC have a lower overall HPV and HR-HPV prevalence than SCC ( $p < 0.001$  and  $p = 0.021$  respectively).

HPV prevalence for types 16 and/or 18 was 76.26%, (CI 71.59% – 80.37%) for SCC and 77.87% (CI 69.72 – 84.32) for ASC+ADC, looking at these two HPV types individually, 238 (66.48%, 61.44 – 71.17) of squamous cervical cancers were HPV 16 positive, whereas ASC+ ADC samples were positive for HPV 16 in 62 (50.82%, CI 42.06 – 59.53) ( $p < 0.01$ ). For SCC samples a total of 45 were positive for HPV 18; 12.75%, (CI 9.53 – 16.41%), comparatively, 43 (34.43%, CI 26.59 – 43.22) of ADC and ASC were positive for HPV 18 ( $p = 0.001$ ) (Table 14).





**Figure 12. HPV type prevalence by HPV type and morphology.** Nonavalent types include HPV 6, 11, 16, 18, 31, 33, 45, 52 and 58. Quadrivalent include: 6, 11, 16 and 18. Bivalent vaccine includes types 16 and 18. Bivalent\* includes types 16, 18 and cross-protected types 31, 33 and 45.

3.4.1.4 Dominant HPV types in cervical cancer samples in Scotland

Dominant types and their proportions for both morphologies are presented in Table 14. HPV 16 was the most dominant type for SCC and ADC+ASC, 238/358 (66.85%, CI 61.80 – 71.54) and 62/122 (50.82%, CI 42.06 – 59.53) respectively. The main difference was the proportion of HPV 18 detected in both groups. HPV 18 had a higher prevalence in the ASC and ADC than in SCC, present in 35.25% (27.34 - 44.06) and 12.64% (9.58 - 16.50) respectively. The detection of 1 case of 31 and absence of type 33 in the ASC+ADC was notable when compare with SCC (5.34 vs 0.82 and 5.06 vs 0 respectively).

**Table 14. High risk HPV types in SCC and ASC + ADC in the cervix.** Data presented as number of times the virus was detected and percentage over the total number of samples for each morphology category. ASC and ADC have been aggregated in one group due to the small number of ASC cases (n=14).

hr-HPV types	SCC		ASC + ADC		Unknown Morphology	
	n	% (CI 95%)	n	% (CI 95%)	n	% (CI 95%)
HPV 16	238	66.85 (61.80 - 71.54)	62	50.82 (42.06 - 59.53)	94	57.67 (49.99 - 64.99)
HPV 18	45	12.64 (9.58 - 16.50)	43	35.25 (27.34 - 44.06)	29	17.80 (12.68 - 24.38)
HPV 31	19	5.34 (3.45 - 8.19)	1	0.82 (0.14 - 4.50)	4	2.45 (0.96 - 6.13)
HPV 33	18	5.06 (3.22 - 7.85)	0	0 (0 - 3.05)	4	2.45 (0.96 - 6.13)
HPV 35	3	0.84 (0.29 - 2.44)	0	0 (0 - 3.05)	0	0 (0 - 2.30)
HPV 39	4	1.12 (0.44 - 2.85)	2	1.64 (0.45 - 5.78)	1	0.61 (0.11 - 3.39)
HPV 45	21	5.90 (3.89 - 8.85)	9	7.38 (3.93 - 13.43)	8	4.91 (2.51 - 9.39)
HPV 51	1	0.28 (0.05 - 1.57)	0	0 (0 - 3.05)	1	0.61 (0.11 - 3.39)
HPV 52	9	2.53 (1.34 - 4.74)	0	0 (0 - 3.05)	1	0.61 (0.11 - 3.39)
HPV 56	5	1.40(0.60 - 3.24)	0	0 (0 - 3.05)	2	1.23 (0.34 - 4.37)
HPV 58	0	0 (0 - 1.07)	1	0.82 (0.14 - 4.50)	3	1.84 (0.63 - 5.27)
HPV 68	4	1.12 (0.44 - 2.85)	0	0 (0 - 3.05)	0	0 (0 - 2.30)

Most dominant types were also obtained for the 3 collection, 2015, 2016 and 2017 (Table 15). Types 16, 18, 45, 31 and 33 remained most dominant types over the years but the order of 3<sup>rd</sup> to 5<sup>th</sup> place changed.

**Table 15. Most prevalent HPV types on all cervical cancer samples (irrespective of morphology) by year.**

Year	5 most common HPV types
2015	16 > 18 > 33 > 45 > 31
2016	16 > 18 > 45 > 31 > 33
2017	16 > 18 > 45 > 31 > 33

3.4.1.5 HPV positivity stratified by types included in vaccines.

Types directly included in the bivalent (16 & 18) and quadrivalent (6, 11, 16 and 18) vaccines were positive in 490 76.44% (95% CI, 73.0 – 79.85) and 494 cervical cancer cases 77.07%; (95% CI 73.66 – 80.16) respectively. When factoring in cross protection (for the bivalent vaccine) a total of 549/641; 85.65%, (95% CI 82.72 – 88.15) samples were positive. Types included within the nonavalent vaccine were positive in 562/641 (87.67%, 95% CI 84.91 – 90.0) of cancer samples. When considering morphology, in SCC, bivalent, quadrivalent and nonavalent types were positive in 76.69% (95% CI 72.03 – 80.78), 77.53% (95% CI 72.91 – 81.56) and 88.76% (95% CI 85.06 – 91.64) respectively. For bivalent types with cross-protection 87.36% (83.50 – 90.42) of SCC were positive. For the combination of ASC+ADC bivalent, bivalent (with cross-protection), quadrivalent and nonavalent types were positive in 77.87% (95% CI 69.72 – 84.32), 82.79% (75.12 – 88.46), 77.87% (95% CI 69.72 – 84.32) and 83.61% (95% CI 76.04 – 89.13) respectively.

3.4.1.6 Trends in HPV positivity during the study period (2015-2017)

Overall HPV prevalence (i.e., any type) on squamous cervical cancer samples has not changed over the 3 years period between 2015 and 2017. HPV positivity rate in 2015 was 93.75% (CI 89.16 – 96.47) while in 2016 and 2017 was 95.54 (CI 89.98 – 98.08) and 94.12 (CI 85.83 – 97.69) respectively. For HR-HPV, positivity was 91.48% (86.42 – 94.77), 92.86% (86.54 – 96.34) and 94.12% (85.83 – 97.69) for 2015, 2016 and 2017 respectively.

Univariate analysis showed no differences in the HPV positivity among the 3 different collection years (p= 0.4582). No differences in positivity were found for HR-HPV types (p=0.469), HPV 16/18 (p=0.124) and

nonavalent types (p=0.940). HPV type specific and aggregated in groups are represented by year in table Appendix 1 and Figure 13.

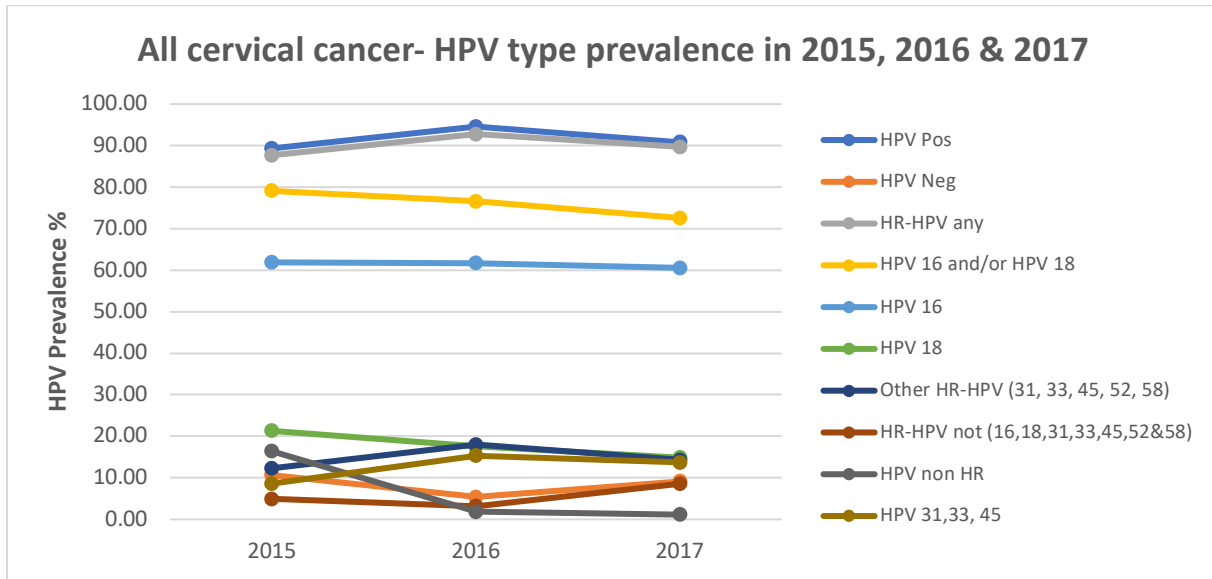


Figure 13. HPV type prevalence in 2015, 2016 and 2017 for all cervical cancer samples.

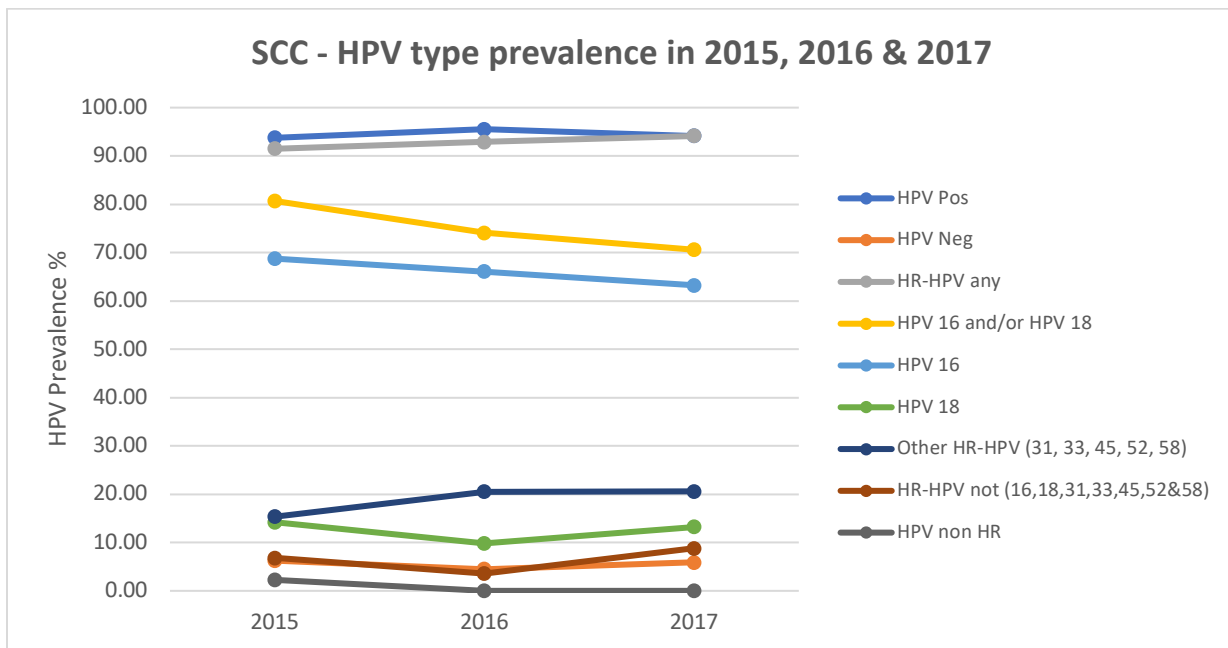
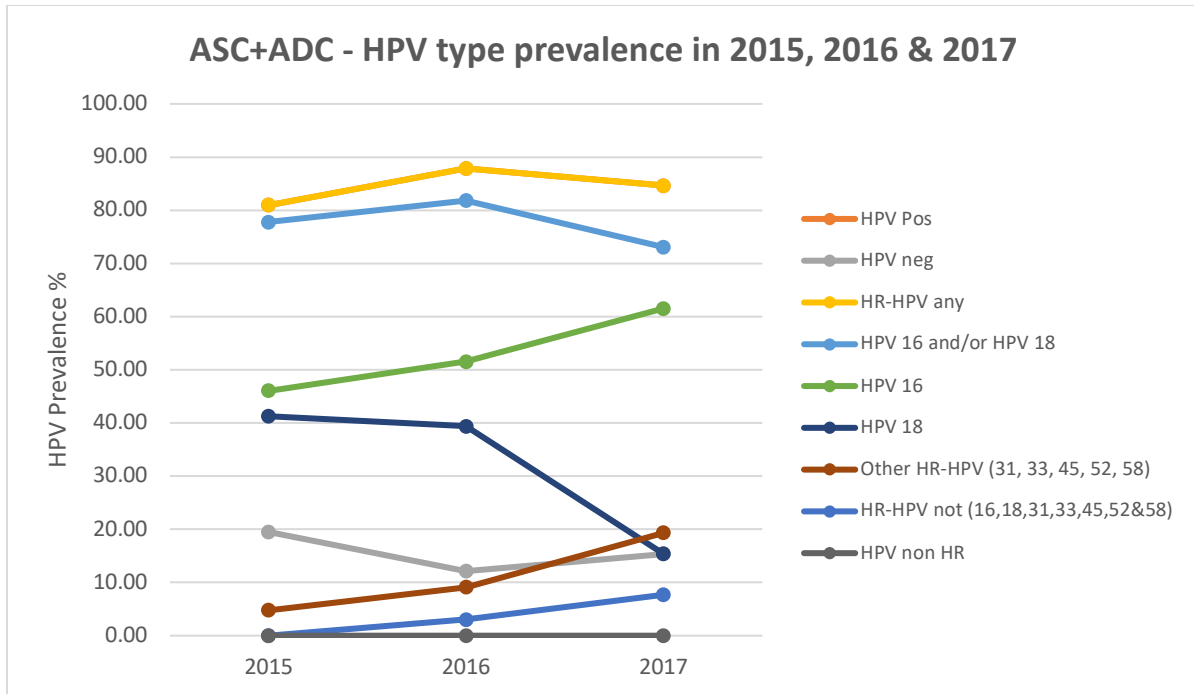


Figure 14. HPV type prevalence in 2015, 2016 and 2017 for SCC samples.



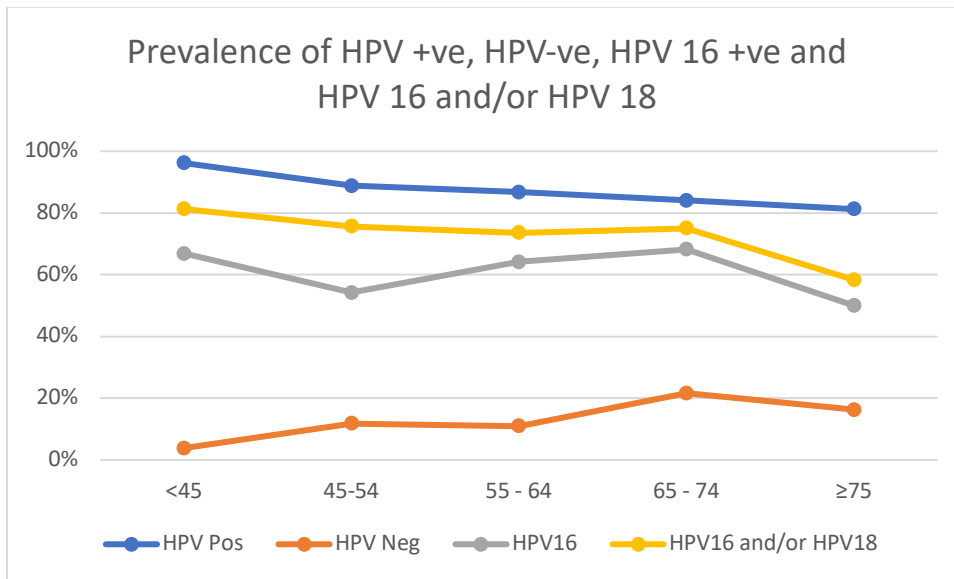
**Figure 15. HPV type positivity in 2015, 2016 and 2017 for ASC + ADC samples.**

### 3.4.1.7 Age and HPV positivity in cervical cancer samples

As detailed in Appendix 1, HPV positivity in cervical cancer was more common in women aged <45 years than >45 years old; 96.17% (95% CI 92.88 – 97.97) and 86.11% (95% CI 81.30 – 89.84) respectively ( $p < 0.001$ ).

When considering detection of “any HPV”, prevalence decreased from 96.17% (95% CI 92.88 – 97.97) in women <45 to 88.79% (95% CI 81.42 – 93.47) in women aged 45 – 54, 86.79% (95% CI 75.16 – 93.45) in women aged 55 – 64 years, 84.09% (95% CI 70.63 – 92.07) in women 65 – 74 and 81.25% (95% CI 68.06 – 89.81) in those aged >75 at diagnosis. When using the <45 y/o group as a reference, adjusted comparison (for age morphology and health board) shows all older groups were less likely to test HPV positive (0.34 (95% CI 0.14 – 0.90),  $p = 0.029$ ).

HPV 16 and or 18 had a higher prevalence in the <45 years old cohort (81.28%, 95% CI 75.80 – 85.75) than in any of the other age groups (appendix 1). However only >75 group 58.33% (95% CI 44.28 - 71.15) had a significant reduced HPV 16/18 prevalence when compared with <45 in the adjusted analysis (p<0.001)



**Figure 16. Changes in HPV positive, HPV-negative, HPV 16-positive and HPV 16 and/or HPV 18-positive prevalence in the different age groups.**

### 3.4.1.8 Does HPV positivity varies by Healthboard location in cervical cancer samples?

The percentage of HPV positive samples on cervical invasive lesions showed a level of variation according to health board of diagnosis (Appendix 1). The NHS health board GGC was used as the reference due to the size of the region and population it covers.

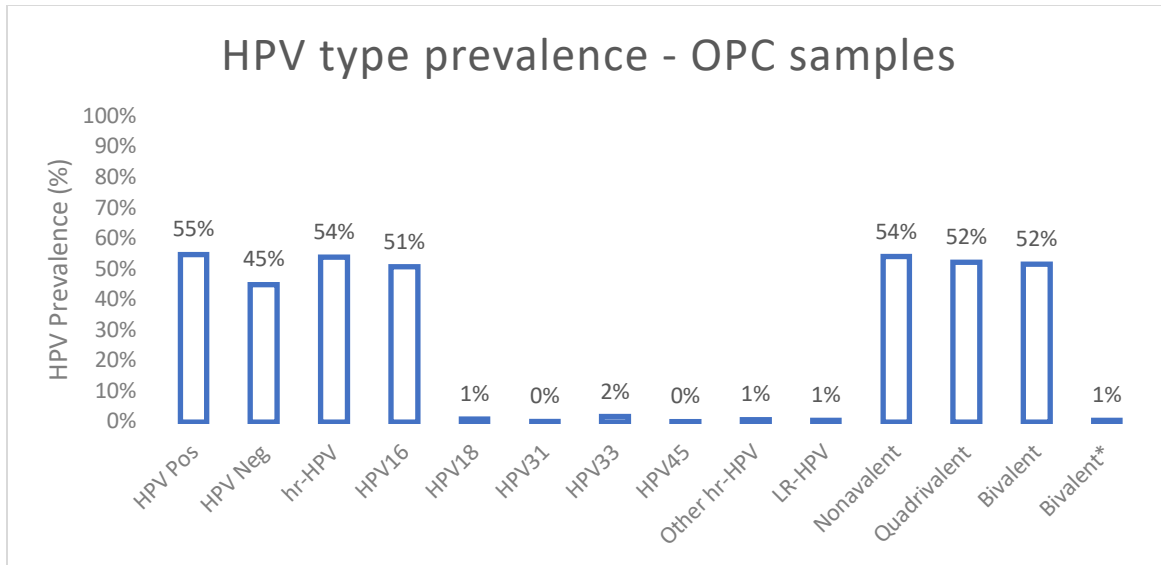
Three NHS health board regions had significant differences in HPV positivity when compared with GGC: Univariate analysis showed Highland (p=0.005), Fife (p=0.045) and Grampian (p=0.003) had the lowest prevalence of HPV with 75.0% (95% CI 55.10 – 88.0), 85.29% (95% CI 75.0 – 92.81) and 76.47% (95% CI

60.0 – 87.56) respectively. After adjustment for age, collection year and morphology, lower prevalence for these 3 regions remained significant. For 16/18 cases, Fife (when GG&C is used as reference) was the only health board with a significant lower HPV prevalence, however when adjusted by age, year and morphology, no significant differences could be identified.

### **3.4.2 HPV type prevalence in oropharyngeal cancer samples**

#### *3.4.2.1. Overall HPV prevalence in the oropharynx*

A total of 1798 oropharyngeal samples were received for HPV testing January 2013 to July 2020 representing 8/14 Scottish health boards. HPV nucleic acid was detected in 987 samples (54.89%; 95%CI 52.58 – 57.18), 811 (45.11%; 95% CI 42.82 – 47.42) being negative. HPV type prevalence is showed in Figure 17 and Table 16. Almost all the HPV types detected were classified as high-risk types (971 (53.57%, 95% CI 51.69 – 56.29). By far, the most prevalent HPV type was HPV 16, detected in 915 samples (50.89%, 95% CI 48.58 – 53.20), 92.71% (95% CI 90.92 – 94.17) of the total positives. HPV 33 was the second HPV type more prevalent with 3.78% (95% 1.26 – 2.50). HPV 18 was the third most common and present in 17 samples (0.95%; 95% CI 0.59 – 1.51).



**Figure 17. HPV type prevalence by HPV type – Oropharyngeal I cancer samples.** Samples have been also aggregated in Other HR-HPVS (35, 39, 51, 56, 59 and 68). Nonavalent types include HPV 6, 11, 16, 18, 31, 33, 45, 52 and 58. Quadrivalent include: 6, 11, 16 and 18. Bivalent vaccine, the first available in the market, includes types 16 and 18.

A total of 12 different HPV types were detected in the oropharyngeal samples either as mono or within multiple infections. A total of 969/1798 (53.89%, 95% CI 51.58 – 56.18) samples were infected with only one type and 18/1798 (1.0%, 95% CI 0.63 – 1.58) samples were positive for at least 2 different types.



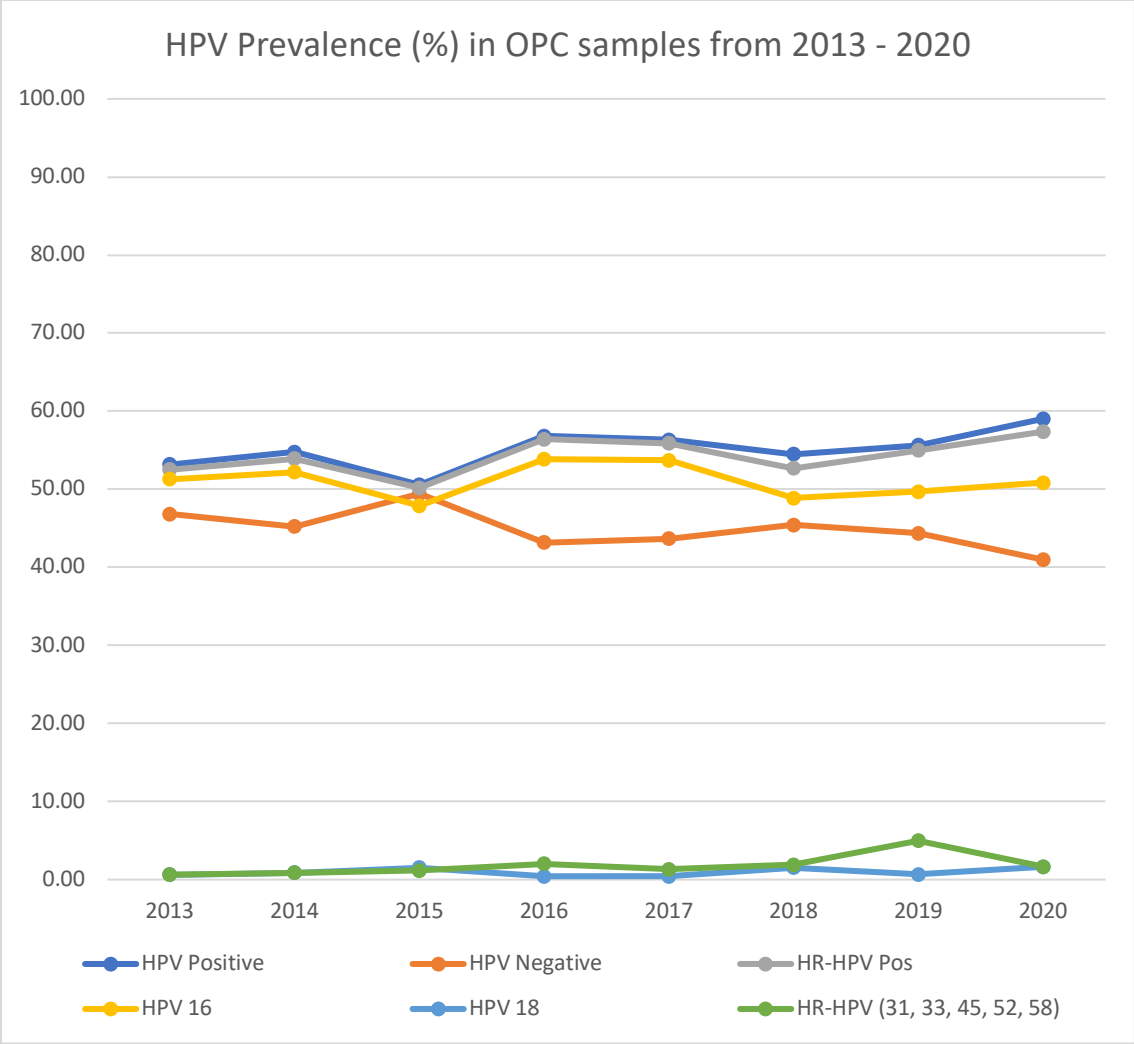
**Table 16. HPV types detected in oropharyngeal cancer samples between 2013 and 2020.**

HPV Type	2013	2014	2015	2016	2017	2018	2019	2020	Total	Percentage N/HPV+ve (95% CI)
HPV 16	81	120	125	126	123	128	150	62	915	92.71 (90.92 - 94.17)
HPV 18	1	2	4	1	1	4	2	2	17	1.72 (1.08 - 2.74)
HPV 31	0	1	1	0	0	1	0	0	3	0.3 (0.1 - 0.88)
HPV 33	1	2	2	4	3	3	13	4	32	3.24 (2.30 - 4.54)
HPV 35	1	0	1	0	1	2	2	1	8	0.81 (0.41 - 1.59)
HPV 39	0	0	0	0	0	0	0	0	0	0 (0 - 0.39)
HPV 45	0	0	0	1	0	0	2	1	4	0.41 (0.16 - 1.04)
HPV 51	0	0	0	0	0	1	1	0	2	0.2 (0.05 - 0.73)
HPV 52	0	0	0	0	0	0	0	0	0	0.25 (0.07 - 0.91)
HPV 56	0	0	0	0	0	1	0	0	1	0.1 (0.01 - 0.57)
HPV 58	0	0	0	0	0	1	0	1	2	0.2 (0.05 - 0.73)
HPV 59	0	0	0	0	0	0	0	0	0	0 (0 - 0.39)
HPV 68	0	0	0	0	0	0	0	0	0	0 (0 - 0.39)

As HPV 16 is the clear dominant type, the preventable fraction of the HPV vaccines for OPC are very similar for the 3 licenced HPV vaccines. Nonavalent, quadrivalent and bivalent (with cross-reactivity) could potentially prevent 54.11% (95% CI 51.86 – 56.46), 52.34%; 95% CI 50.03 – 54.64), 53.35% (95% CI 51.19 – 55.80) of OPC respectively.

#### 3.4.2.2 Does HPV positivity change over time? – Oropharyngeal cancer samples

Overall HPV prevalence on oropharyngeal cancer samples has not varied over time. No differences were found on HPV positivity rate among the different collection years ( $p = 0.278$ ). Figure 18 represents the extent of HPV positivity among the 8 years of testing.

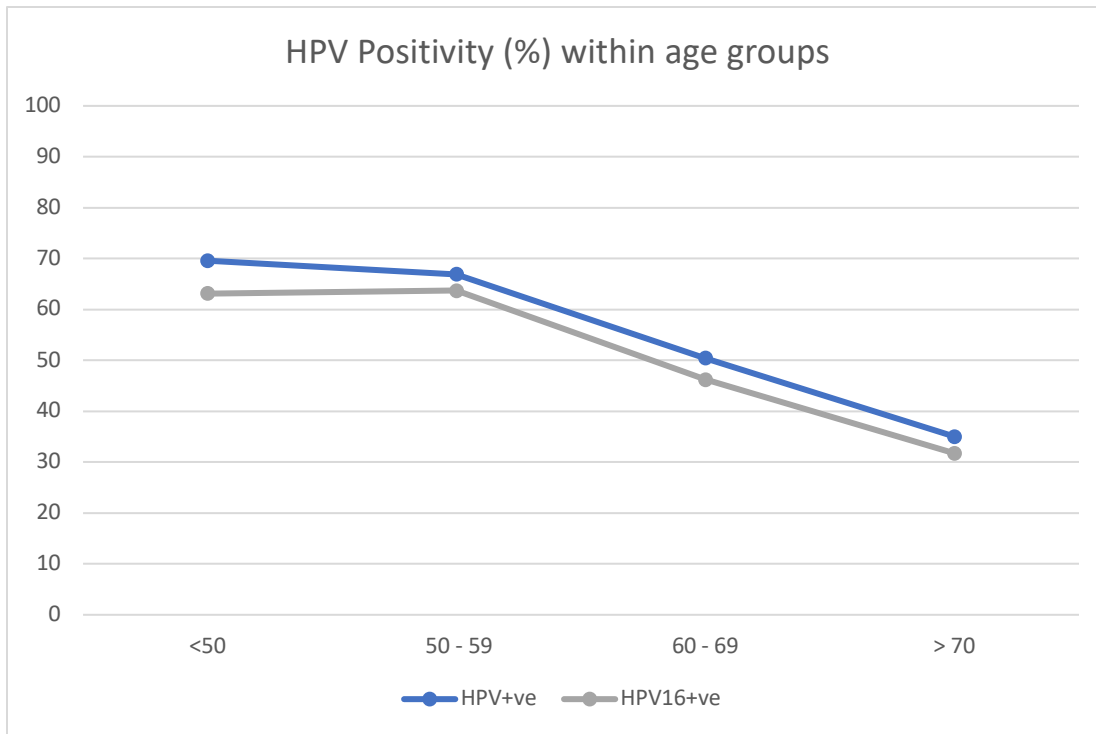


**Figure 18. HPV type prevalence in oropharyngeal cancer samples from 2013 to 2020.**

*3.4.2.3 HPV positivity by age – Oropharyngeal cancer*

Overall HPV status (positive and negative) decreased over age ( $p < 0.001$ ). Figure 19 and Appendix 1 represents positivity stratified by age overall HPV and for HPV 16. Overall HPV positivity was 69.59% in those <50 years old, 66.95% in 50-59, 50.42% in 60-69 and 35.03% in <sup>3</sup>70. For HPV 16/18 positivity, these figures were 63.13% in <50, 63.73% in 50-59, 46.21% in 60-69 and 31.73% in >70.

Logistic regression analysis showed a significant reduction of HPV positivity in age groups 60-69 and >70 in the univariate and adjusted analysis when used <50 as reference ( $p < 0.001$ ). The same reduction was recorded in the same age groups for HPV 16-positive cohort in the univariate and adjusted analysis ( $p < 0.001$ ). More details are described in Appendix 2.



**Figure 19. HPV positivity by age group for oropharyngeal cancer samples.**

#### 3.4.2.4 Does HPV positivity vary by sex in oropharyngeal cancer?

Overall HPV prevalence was higher in men (58.25%, CI 55.20 – 61.24) than women (44.89%, CI 40.49 – 49.37) (OR 1.68,  $p < 0.001$ ). The proportion of HPV 16 was also higher in men (54.81%, CI 52.11 – 57.48) than women (40.08%, CI 35.79 – 44.53) (OR 1.63,  $p < 0.001$ ).

### 3.5. Discussion

This chapter presented detailed information on the type specific diversity and prevalence of HPV in the most common HPV driven cancers in Scotland: (cervical and oropharynx). These data provide information

into the percentage of cases that could be prevented in the future through vaccination and also and give insight into the extent disease that is unlikely to be reduced by vaccination.

In the cervical cancer cohort, HPV was detected in 91.58% of the cervical cancer samples, with 90.02% associated with high-risk types. In SCC samples, 94.38% were positive for any HPV type, while 91.90% ASC+ADC had a lower HPV prevalence, with 83.61% of samples positive for any HPV type.

One of the most comprehensive studies looking at HPV type specific prevalence in different world location, De Martel *et al.*<sup>63</sup>, found that the relative contribution of HPV 6, 11, 16, 18, 31, 33, 45, 52 and 58 to the world burden of cervical cancer was 89.5% (470,000 cases, 2012) Moreover, local comparison (Scotland) with data published by Cuschieri *et al.*<sup>68</sup>, from samples collected prior to 2004, shows that overall HPV positivity has increased only marginally from 88% to 92% and HPV 16/18 from 72% to 75%. Most of country-specific analyses of cancer are cross-sectional so the fact that cancers in Scotland we have surveyed cancers at different time points (2015, 2016 and 2017) indicates the stability of the HPV associated component which is beneficial when considering vaccine impact. Another study performed in the UK, Mesher *et al.*, merged data from the 4 UK nations, and found that overall HPV prevalence in ICC was 95.8%, prior to HPV immunisation.<sup>108</sup> Although, they found significant differences in terms of hr-HPV prevalence between the UK countries, Scotland was the country with the lower prevalence (83.2%) while Wales had the highest prevalence (97.6%). This chapter has found that overall cervical cancer had a prevalence of hr-HPV types of 90%, 83.6% for SCC only.

In terms of non-HPV cases, a total of 8.42% of the cervical cases did not test positive for any HPV by the PCR based tested used in the Scottish reference lab. By identifying those HPV negative cases, we are now capable of assessing the scope and characteristics of this component. Gaining a deeper understanding of the HPV-negative elements of these cancers could allow us to customize future prevention and treatment strategies and maybe adapting the screening program for these HPV negative cases by performing extra tests. A recent publication from Arroyo Mühr *et al.*, 2020<sup>62</sup> showed that 43.11% of HPV-negative cervical

cancers (by PCR) had HPV detected after using next generation sequencing (NGS) and that the majority of these were positive for high-risk or probably high-risk HPV types. This suggests, it is feasible that some types have not been detected due to the molecular assay choice/sensitivity or possible partial or complete missing of the target region due to integration.

In the oropharyngeal cohort, HPV prevalence identified in the oropharynx cohort was 54.89%. High-risk types were identified in 53.57% of the cases. This aligns with prevalence obtained in other national studies. Schache *et al.*,<sup>109</sup> identified that overall proportion of HPV-positive OPSCC between 2002 and 2011 in the UK was 51.8%. Similarly, Wakeham<sup>107</sup> *et al.* found that in the Scottish sample cohort, HPV was detected in 60% of cases while Wells *et al.* found that HPV was present in 57% of samples. Moreover, dominance of HPV 16 (92.71% of OPC) aligns with what previously identified in the 3 publications, where HPV 16 was present in 96.3%<sup>109</sup> 90%<sup>106</sup>, 93.38%<sup>107</sup> of HPV-positive OPC respectively.

The analysis of HPV prevalence and its association with different demographics has provided an understanding of the differences in HPV prevalence between ages, regions of Scotland and changes over time. For those cervical cancer, samples where age information was available, it has been identified that HPV positivity declines with age ( $p < 0.05$ ). The highest positivity for HPV was detected in women aged < 45 years old (96.17%) decreasing in the older groups to 81.25% in women aged  $\geq 75$  years old. This also applies to HPV 16/18 positivity, where significant differences between  $\geq 75$  years old group (58.33%) and <45 years (81.28%) was shown. The reduction of HPV 16/18 with age seen in the collected cervical cancers is consistent with previous publications<sup>110,111</sup>. The reasons for the reduction of HPV prevalence with age and increase of non-HPV cervical cancers, are not fully understood. It is plausible that there may be a greater opportunity for the HPV to be “lost” during the carcinogenic process in the elderly<sup>112</sup> or maybe with aging there is a greater chance that non-HPV cancerous changes/pathways may play a larger role<sup>113</sup>.

When looked at the differences in cases of CCX and ADC+ASC, there was a greater proportion of glandular cases in the 60–69-year group than in the CCX cohort. Moreover, HPV positivity decreases dramatically in the ADC+ASC cases with age, from 90% in the <45 years to 72% in the 55-74 and 37.5% in the >75. Again, the reasons for this reduction and not fully understood.

In the OPC cohort, prevalence of HPV positive cases tends to be higher in <60 years old than in the older population. Overall HPV positivity was 69.59% in those <50 years old, decreasing to 50.42% in 60-69 and 35.03% in >70. This is associated to the fact higher prevalence of non-HPV OPC increase with age. Older patients are more likely to have drunk and smoked heavily, both important risk factors<sup>114</sup>, more common in older men, with a median age of 61<sup>115</sup>.

When HPV positivity in cervical cancers was analysed taking into consideration the healthboards, data obtained suggested that Fife, Grampian and Highland were the locations with the lowest HPV prevalence. In terms of HPV 16/18, no differences could be found in the univariate and adjusted logistic analysis, suggesting a consistent HPV 16/18 prevalence across the country.

It was not possible to perform the same healthboard analysis for the oropharyngeal cohort, as not all the healthboards sent all the oropharyngeal cases to the reference lab for testing.

Currently, the HPV vaccine offered in Scotland and the United Kingdom, is Gardasil 9 (started in 2021/22 school year). When considering the types present in the Gardasil-9 for the cervical cohort, it could potentially prevent a total of 87.67% of cases, 10% above the other bivalent and quadrivalent vaccines. When factoring in cross protection (for the bivalent vaccine) a total of 85.65% samples could be potentially prevented. For the OPC cohort, a total of 52-54% could potentially be prevented with any of the vaccines.

It is interesting to reflect on the fact that in terms of cancer prevention the difference between the nonavalent vaccine and the bivalent (when considering cross-reaction) vaccine is relatively small. This is consistent with a Cochrane review, where it was described, that bivalent vaccine offered more complete protection against cervical pre-cancer overall than the quadrivalent vaccine<sup>116</sup>.

The investigations presented in this chapter have some limitations and challenges. One of them is the number of cases and NHS health board distribution of the oropharyngeal cancers. Even if the genotyping service offered by the reference laboratory was free of cost for every NHS healthboard, not all used the service consistently therefore not all oropharyngeal diagnosed cancers were tested. However, the largest healthboard in Scotland (NHS GGC) sent all the oropharyngeal cases. Using Glasgow's data, we can extrapolate and obtain a general view of the prevalence in Scotland.

Moreover, there is some age and histology data missing. ISD was contacted to retrieve the missing information from their databases, but by the time this chapter was prepared no information was received. Covid have also had an impact in ISD, and resources were focuses in covid data analysis. Therefore, the actual analysis of age and histology was incomplete due to the missing data.

Data analysis within this chapter has allowed to identify that positivity of some HPV types have decreased over the years in the cervical cohort. However, the positivity of HPV in cervical and oropharyngeal cancer lesions has not changed over time.

The data analysed within this chapter provides contemporary information on epidemiology in the two most common HPV driven cancers in Scotland, showing that some HPV types have decreased over the years in the cervical cohort, but no changes have been identified in the OPC samples. It also highlighted the proportion of cases that are HPV negative. By previous studies, we know that HPV negative cervical

cancers do worse than those where HPV has been detected. As screening is now based in HPV status, it could potentially be good to revise and consider changes in the program for a better tailored management. Maybe an extra genomic test for those HPV negative to discard completely the presence of HPV in the cancer. However, one of the limitations would be access to sequencing technology and the cost associated as it could result excessive for a population-based programme.

In the oropharyngeal cohort, it will take another 10-15 years to see the full effect of the HPV vaccine on HPV positive OPC. However, as almost 50% of the OPC in Scotland are not associated with HPV, it seems necessary to investigate about other routes for the OPC prevention. And although number of OPC have increased in Scotland in the 7 years of samples analysed, this is related to the increasing of both HPV positive and negative component. This aligns with what has seen in the 4 nations of the UK. However, HPV positivity in OPC must continue to be monitorised as the USA has registered an increase of HPV positive OPC<sup>117,118</sup>.

Data collected from the screening and HPV testing programs put in place in Scotland for cervical and oropharyngeal cancer have allowed to obtained HPV status and type information that can be used to help understanding what the main types associated with cervical and oropharyngeal cancer are, identify if there are any specific association of HPV infection with demographics and if there are any specific regions of Scotland with a higher or lower incidence. In contrast, HPV positivity in anal cancer is not available in Scotland due to a lack of research in this area and the absence of a screening infrastructure that would support surveillance.



## **4.Role and influence of HPV in anal disease in the South-East of Scotland. HPV type specific prevalence and viral load.**

*Please note that part of the results presented in this chapter have been published: Guerendiain D et al., (2022). HPV status and HPV 16 viral load in anal cancer and its association with clinical outcome.*

### **4.1 Introduction**

As described in the previous chapter, HPV was detected in 92% and 55% of cervical and oropharyngeal cancers respectively genotyped for HPV in Scotland. Comparatively there is a lack of data on anal cancer. Understanding the epidemiology of HPV in anal cancer lesions will help determine which HPVs are prevalent in anal disease in Scotland. Linking this data to follow up information also permits the opportunity of assessing the influence of HPV status on clinical outcome. Given the morbidity of this disease, understanding what component of disease is vaccine-preventable and whether aspects of HPV infection could be used for risk stratification in the affected/disease population is important.

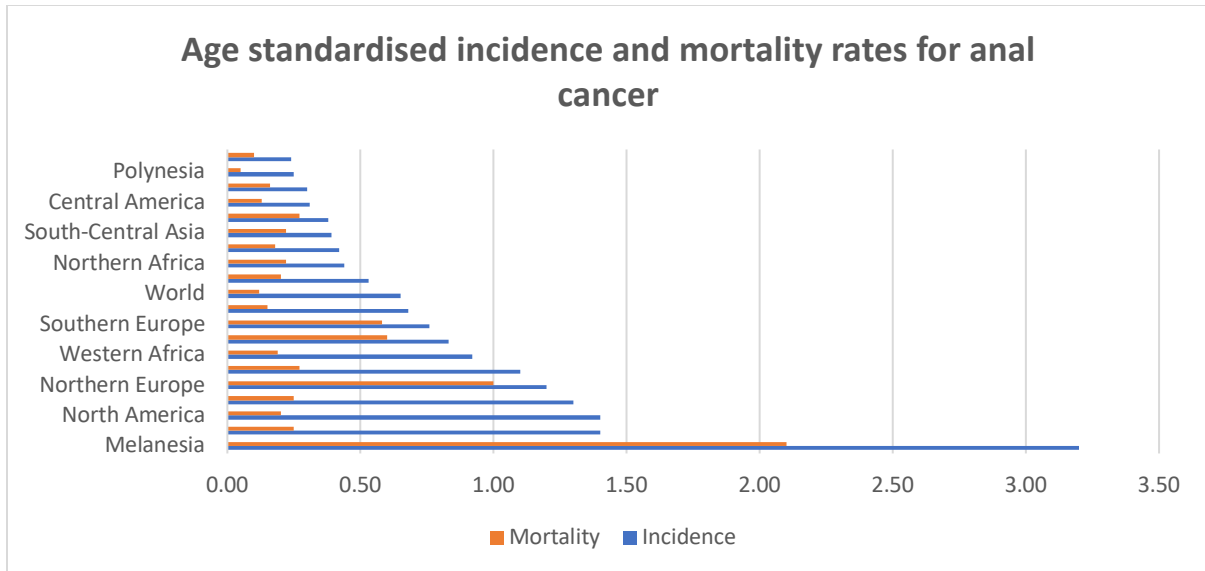
#### *4.1.1 Anal disease*

Anal lesions appear in the skin around the back passage (perianal skin), but they can also be found inside the anal canal. As in the cervix, HPV can drive low-grade and high-grade lesions, which can lead into cancer. These lesions are denominated AIN and are classified in 3 different grades according to the extent of histological abnormality (AIN1, 2 and 3) <sup>119,120</sup>; similar to the cervix, abnormal cells can clear on their own. While the natural history of AIN is not as well defined as CIN, AIN 2 and 3 are considered high grade lesions and may be treated if detected <sup>120</sup>. Invasive lesions cannot reverse, and treatment will be required (surgery, radiotherapy, chemotherapy, or combined therapy).

The risk of anal cancer can be heightened by various factors such as age (with a greater incidence observed in those aged 50 or above), multiple sexual partners, engaging in anal sexual activity, smoking, prior history of cervical cancer, presence of HPV and underlying immune system disorders<sup>121–126</sup>. Women who have or had high-grade lesions or worse in the cervix have a higher risk of anal lesions and cancer when compare with HPV-negative women<sup>127,128</sup>. Men who have sex with men (MSM) also have higher risk of anal cancer. Additionally, people living with HIV<sup>129,130</sup> have a higher risk of anal cancer; incidence of anal cancer is 40 to 80 times higher in the population living with HIV<sup>121–126</sup>.

#### *4.1.2 Global epidemiology – Anal cancer*

Approximately 40,000 cases are diagnosed per year worldwide, from which, 31,600 are theoretically preventable<sup>131</sup> through vaccination programs. The World Health Organization published the incidence and mortality data of anal cancer in 2018. The global incidence for anal cancer is 0.53 age-standardised rate (ASR) while the highest has been recorded in Melanesia and the lowest in Eastern Asia. Northern Europe has an incidence of 1.1.<sup>132</sup> Looking at ASR mortality, the world rate is 0.20 while Melanesia and Middle Africa have the highest mortality (0.77 and 0.85 respectively) and Central America the lowest mortality (0.05)<sup>132</sup>. Northern Europe has the higher mortality rate among the European regions (0.31 vs 0.15 and 0.21). When compare with the ASR incidence of cervical cancer (ASR 13.1), and oropharyngeal cancer (ASR1.1), anal cancer incidence remains much lower (ASR 0.5)<sup>104,132,133</sup>.



**Figure 20. Age Standardized per 100,000 (ASR) incidence and mortality rates for anal cancer worldwide and by regions.** Source Globocan 2018<sup>14</sup>.

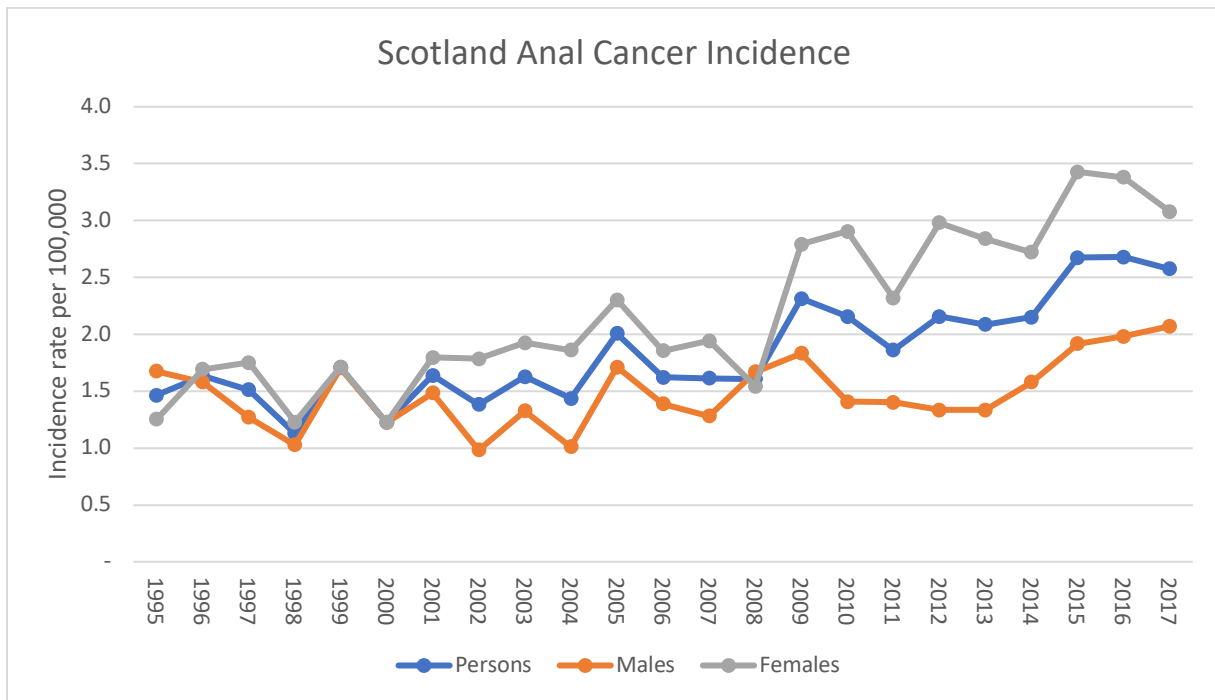
As with other HPV-driven cancers, anal cancer prevalence is increasing worldwide, including USA and Europe<sup>131,132</sup>. Islami *et al.* (2017) looked at the trends in anal cancer incidence rates in 18 countries<sup>134</sup>. They found that anal squamous cell carcinomas (ASCC) incidence rates increased for both sexes in all countries except for India, Estonia, Japan, Singapore and Spain. In England, Robinson *et al.* (2009) looked at the incidence of HPV-related anal cancer in the Southeast England<sup>135</sup> and found an increase in age-standardised rates for both men and women. In men, ASR increased from 0.79 in 1960 – 64 to 1.06 per 100,000 in 2000 – 2004, while in women, the rate of increase was higher, rising from 0.45 per 100,000 people during 1960-64 to 1.18 per 100,000 people between 2000-2004.

#### 4.1.3 Anal cancer in Scotland

In Scotland, anal cancer data is available through ISD which incorporates the cancer registry and, which contains population data from 1972. Using this data, different investigators have looked at changes in the number of anal cancer cases over time (Figure 21). Brewster *et al.* (2016) analysed the incidence of

squamous cell carcinoma in the anus in Scotland between 1975 and 2002. They found that over the 27 years, the risk of squamous cell carcinoma of the anus in Scotland had more than doubled in both sexes<sup>136</sup>. In 2014, Wakeham *et al.* also used data from ISD to investigate the anal cancer incidence from the period between 1972 to 2011<sup>137</sup>. They found numbers of cases increased from 97 in the 1972 – 1976 period to 476 in 2007-2011, a 390% increase between both dates. Looking at the incidence (EASR per 100,000), they also found an increase from 0.6 to 1.9 (increase of 217%).

According to most recent data obtained from ISD (produced on request of the author), there has been a 120% increase from 1995 to 2017 in Scotland<sup>64</sup>. The biggest increase has been noted in the 55-59 (2.2 to 8.2) and 65-69 (1.7 to 7.5) ages groups.



**Figure 21. European Age-Standardised Incidence Rates per 100,000 population, Scotland.** Source ISD Scotland. Accessed March 2021 <sup>21</sup>.

#### 4.1.4 HPV epidemiology in anal diseases

Similar to cervical cancer, most anal cancers (~90%) are caused by persistent infections with high-risk HPV types<sup>119,120</sup>, HPV 16 being the most prevalent<sup>138</sup>. De Sanjosé *et al.* (2019) performed a cross sectional study looking at HPV-related cancers. They collected 496 anal cancer cases from 50 countries (although no anal samples had been collected from the UK) which were subsequently analysed for HPV. According to their publication 95.3% of the anal cancers were positive for HPV, from which 84.3% were positive for types 16 and 18 (80.7% HPV 16)<sup>131</sup>.

A recent review and meta-analysis on prognostic significance of HPV DNA and p16INK4A in anal cancer found that patients positive for HPV or p16INK4A positive may have a better overall survival compared with HPV DNA or p16INK4A negative<sup>139</sup> cases. Two studies were performed in the UK in 2013 and 2015 where they looked at prognosis of HPV-driven cancer. Baricevic *et al.* (2015) found that HPV 16 positivity was significantly correlated with improved 5-year relapse free and overall survival<sup>140</sup>. Gilbert *et al.* (2013) studied 153 patients with anal cancer and found p16INK4A negative patients had significantly worse overall survival<sup>141</sup>. Improved survival in HPV-driven anal cancer reconciles with what has been observed in other HPV-driven cancers as in the cervix, vulva, head and neck and penis<sup>142-146</sup>. However, for anal cancer, studies performed in the UK have so far only analysed a relatively small number of patients. This work will include the biggest single cohort of anal lesions tested for HPV DNA in the Scotland and the whole UK.

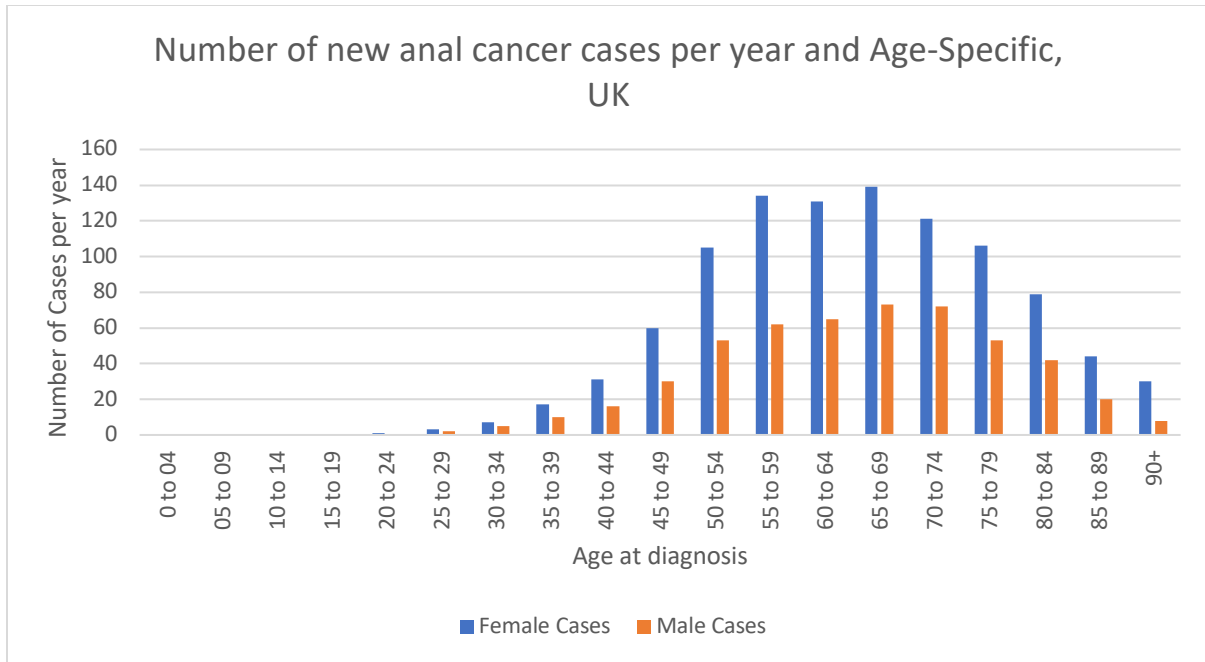
#### 4.1.5 Epidemiology in anal disease in the UK/Scotland

Although anal cancer incidence has increased in in the last years, no data has been published about HPV type specific prevalence in high-grade anal lesions and anal cancer in the UK. There is some data on the prevalence of HPV in residual rectal swabs from asymptomatic men attending GUM, but this has largely

been confined to MSM. According to Cameron *et al.*, 2020 HPV was detected in 72.8% (95% CI 70.2% to 75.3%) of anal swabs, with at least one high-risk type being present in 59.1% (95% CI 56.3% to 61.9%) of the samples<sup>65</sup>. In 2015, King *et al.* performed a similar study in MSM and found that 64.9% of anal swab samples were positive for any HPV type and 38.6% for any nonavalent vaccine type (6, 11, 16, 18, 31, 33, 45, 52 and 58)<sup>147</sup>.

When considering the impact of primary prevention on anal disease, the implementation of the HPV vaccine in boys and girls could clearly exert a significant impact in the reduction of HPV driven anal lesions. According to Cancer Research UK, the largest number of anal cancer cases are registered between the age of 65 and 69, with 14% of the cases (Figure 22). Therefore, it would take some time to see the full effect of the HPV vaccination program on the anal cancer incidence. More information about vaccination in Scotland can be found in the following link:

<https://www.nhsinform.scot/healthy-living/immunisation/vaccines/hpv-vaccine>.



**Figure 22. Number of new cases of anal cancer by age-specific group in per year in the UK (average data from 2016 – 2018).** Figure prepared from data obtained from Cancer Research UK<sup>32</sup>.

#### 4.1.6 Implications of viral load in anal lesions

Various studies have evaluated the HPV viral load in different cancers caused by HPV and its correlation with overall survival, prognosis, and/or as a biomarker for the progression of lesions. Investigators have looked at the association of prognosis and viral load in cervical cancers<sup>148,149</sup> and oropharyngeal cancers<sup>150–155</sup> finding that high viral load is associated with a better prognosis than those with low viral load. A small number of publications exist on the implications of HPV viral load in anal cancer<sup>156–160</sup>, however, the majority have investigated HPV load only in the HIV positive cohort<sup>157–159</sup>, using standard PCR based approaches. The studies revealed that patients who had low median viral load of HPV 16 DNA and low p16 expression had considerably poorer local control and overall survival compared to those who had higher median viral load<sup>160,161</sup>.

The exact reason for why a high HPV viral load is associated with a better prognosis or overall survival is not fully comprehended, although it has been posited that cancer samples with lower HPV viral load could be a result of HPV integration. This integration could potentially be disrupting the repression of E1 and E2 genes, resulting in an increased expression of oncoproteins E6 and E7 but affecting the normal replication of the virus, leading to low number of HPV copies<sup>162,163</sup>.

#### *4.1.7 Current management of anal disease in Scotland*

The morbidity linked with treatments for anal cancer can be substantial. The treatment options include chemoradiotherapy (CRT) or surgical removal of the affected tissue, depending on the location of the cancer (tumors located at the anal margin are treated differently than those in the anal canal)<sup>164,165</sup>. Treatment for anal cancer can significantly diminish the patient's quality of life, with reported morbidities including issues with sexual function and faecal continence<sup>166–168</sup>.

In Scotland, there is no anal screening programme in place at the time this thesis (early 2023) and therefore, anal lesions can only be detected once they are symptomatic. Also, routine HPV testing is not performed on resected AIN or cancers. Therefore, there is a lack of contemporary data on the extent and implications of HPV in anal disease in Scotland.

#### *4.1.8 Chapter Aim:*

The aim of this chapter is to provide information on HPV prevalence in anal cancers collected in the east of Scotland during a 10-year period, by performing a molecular assessment of a cohort of anal cancers to determine specific prevalence and to assess the implications of HPV status and HPV viral load on clinical outcomes including survival.

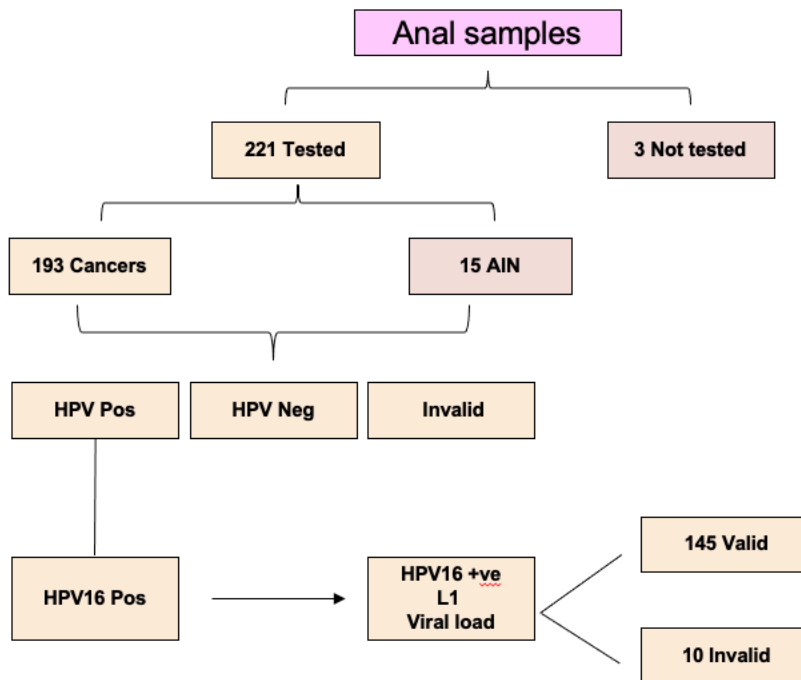


## 4.2 Material and Methods:

### 4.2.1 Demographic and clinical data collation

A total of 221/224 anal lesions were tested for HPV using a PCR-based genotyping test (Seegene Anyplex II HPV28). Samples of anal lesions were collected as part of the standard care of patients with anal disease diagnosed between 2009 and 2018, and preserved in formalin-fixed paraffin-embedded (FFPE) blocks.

Figure 23 shows the characterization process flow to obtain HPV status (positive or negative), type information and viral load for sample in this chapter.



**Figure 23. Overview of the process followed including HPV genotyping of 221 anal samples, using the Seegene Anyplex II HPV28 and viral load analysis of all HPV 16-positive cancer samples through L1 ddPCR.**

Clinico-demographic information such as age, sex, cancer stage (using the American Joint Committee on Cancer (AJCC) TNM system), response to treatment, date of diagnosis, and vital status (dead/alive) were

obtained. Information was collected in January 2020, indexed with a study number, and censored in July 2020 for vital status information and date of death. The cohort was followed from the date of diagnosis until death or date of censoring.

Cases categorized according to the various clinical and demographic variables are summarized in Table 17. Age was stratified in 4 different groups: <50, 50-59, 60-69 and >=70. Lesions were classified as high-grade and invasive. Recurrence was organized in 3 groups: yes, no, or unknown following the ESMO guidelines for anal cancer<sup>164</sup>. Cancer stage was aggregated in 5 groups: 0, I, II, III, IV following AJCC system effective January 2018<sup>45</sup>.

**Table 17. Demographic & clinical characteristics of the anal samples collected between 2009 to 2018 in the South-East of Scotland.** N correspond to the total number of samples for every category described above. Percentage (%) was calculated from the total number of valid samples (208). <sup>a</sup>Due to the small number of samples, stage of cancer was aggregated in I, II, III and IV for the survival analysis on cancer samples. <sup>b</sup>Recurrence (yes) includes recurrence, residual and progression, while “no” includes remission cases. For the anal cancer cohort, response to treatment was classified as yes, no, or unknown and vital status as alive or deceased.

Variable	Level	n =	%
Sex	Female	135	64.90%
	Male	73	35.10%
Age	<50	35	16.83%
	50 - 59	58	27.88%
	60 - 69	58	27.88%
	70 and over	57	27.40%
Type of Lesion	High-grade	15	7.21%
	Invasive	193	92.79%
Collection year	2009	18	8.65%
	2010	13	6.25%
	2011	16	7.69%
	2012	17	8.17%
	2013	18	8.65%
	2014	23	11.06%
	2015	28	13.46%
	2016	26	12.50%
	2017	25	12.02%
	2018	24	11.54%
Stage <sup>a</sup>	0	21	10.1%
	I	29	13.94%
	II	66	31.73%
	III	55	26.44%
	IV	35	16.83%
Recurrence <sup>b</sup>	Yes	17	8.17%
	No	139	66.83%
	Unknown	52	25.00%

#### *4.2.2 Approach to HPV type annotation and viral load quantification*

A total of 221 anal cancer samples were annotated for HPV type-specific prevalence using the Anyplex II 28 assay (Seegene, Korea) centrally at the Scottish HPV Reference Laboratory, Edinburgh, UK. One 10 µm section per FFPE block was obtained using a Leica microtome. DNA extraction was performed using the Microlab Nimbus IVD with the StarMag Universal Cartridge Kit. Genotyping was done with the Anyplex II HPV28 (Seegene) kit.

The analysis of viral load was limited to samples that tested positive for HPV 16, either as a mono-infection or as part of a mixed infection. This was because HPV 16 was the predominant type detected in the study cohort. DNA was extracted using the Qiagen DNA mini kit, and tested at the Centre for Virus Research, University of Glasgow. Viral load was obtained as described in Stevenson *et al.*, 2020<sup>169</sup>. For each sample, the viral load was determined by measuring it against the endogenous RRP30 cellular gene internal control, which had two copies per cell. After retesting, consistent invalids were excluded from the analysis (n=9). The individual HPV 16 viral loads were ranked from smallest to largest and separated using tertiles in three groups: low, medium and high viral load.

#### *4.2.3 Analysis of HPV positivity and viral load influencing survival in anal cancer samples.*

Analysis of HPV positivity and viral load and influence on survival was only performed in the cancer group. Association between HPV overall and HPV type with overall survival was calculated by using Kaplan Meier model. Recurrence was defined as observable and diagnosable signs or symptoms after treatment was received by the patient.

To determine if HPV related anal cancer lesions have better survival (vs HPV negative lesions) a survival analysis was performed. In this case, a Kaplan Meier curve was produced stratified by HPV status (positive/

negative) and also HPV 16/18 versus negative cases and the results of a log-rank test were presented. Survival time is number of days from baseline (define what that is) to date of death or where no date of death is available the patient is considered to be alive and censored on the date of data linkage with TRAK (06 August 2020). Cox proportional hazard ratio model was used to determine the how covariates interacted with survival. Univariate and multivariate analysis were performed.

The univariate and multivariate hazard ratios of HPV status (negative vs. positive) and virus load (low vs. median and high) for all cause death were derived using Cox proportional hazard model. Two multivariate models were derived – age (<50, 50-59, 60-69, 70+) and sex adjusted model and fully adjusted model, where age, sex, stage (I, II, III, IV) and response to treatment (no, yes) were adjusted for. All the statistical analysis were performed using R-studio (version 1.2.1335)<sup>214</sup>. Association of viral load with sex and stage cancer was calculated using Pearson's Chi-squared test.

## **4.3 Results**

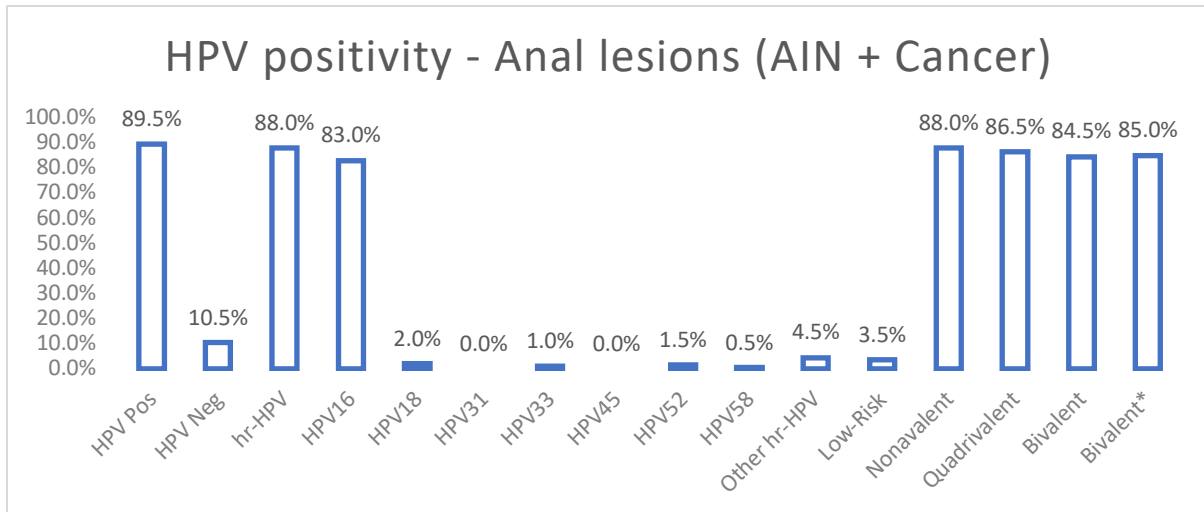
### *4.3.1 Final study samples and exclusions*

A total of 224 samples were received from NRS Bioresource (SR1283), of these it was not possible to extract 3 due to operational issues (instruments were prioritised for COVID-19 testing at the time). Of the remaining 221 anal cancer samples, 13 (5.88%) samples were invalid (on repeat) due to endogenous control failure. A further, 8 anal cancer samples were not included due the absence of clinical details. From the total number of valid samples (200), 193 (92.50%) were classified as anal cancer and 15 (7.50%) were classified as Anal intraepithelial neoplasia grade 3 (AIN3) (Figure 23).

### *4.3.2 HPV type specific positivity in a cohort of 200 High grade Lesions and Cancers in Scotland, diagnosed between 2009 and 2018.*

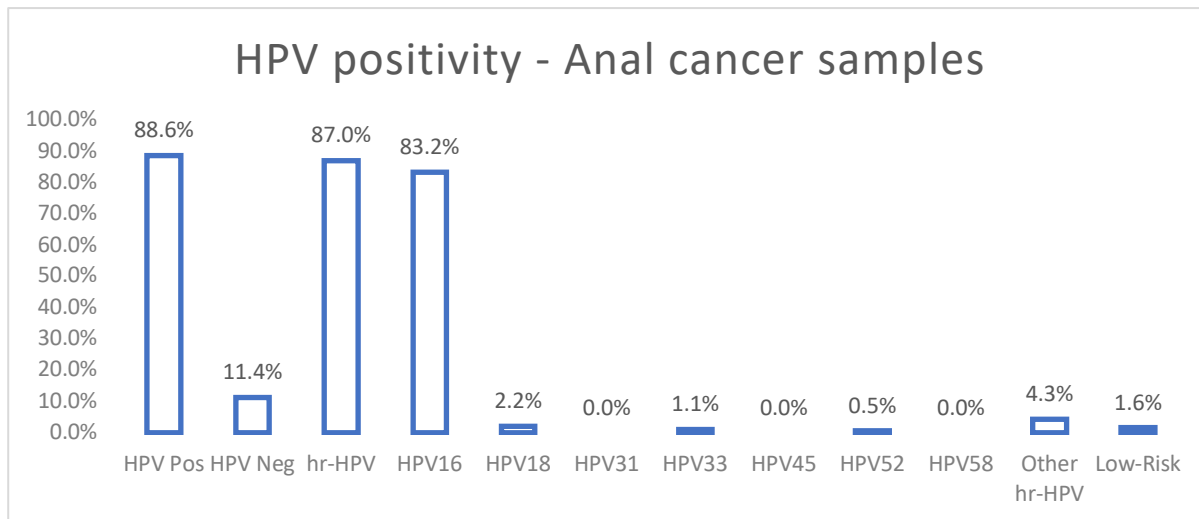
A total of 200 samples had a valid result from which 179 (89.50%, 95% CI 84.48 – 93.03) were positive for HPV and 21 (10.1% 6.70 – 14.95) HPV negative. From the 179 HPV positive samples, 12 had multiple HPV types detected (6.42%) while the rest (167) were mono-infections.

Figure 24 describes the number of HPV types detected in the anal lesions. The majority of HPV types detected were high-risk (98.32%, 95% CI 95.18 – 99.43) of which HPV 16 was the most prevalent (92.74.05%, CI 87.98 – 95.71) followed by HPV 6 (2.79%, CI 1.2 – 6.37).



**Figure 24. HPV Prevalence in 200 anal lesions (AIN + cancer).** HPV Prevalence in 200 anal lesions (AIN + cancer). Types 16, 18, 31, 33, 45, 52 and 58 were discerned individually. High-risk group includes: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68, other-hr-HPV types group includes 35, 39, 51, 56, 59 and 68 and low-risk HPV types 6, 11, 40, 42, 43, 54, 44 and 61). “Bivalent types represent: 16, and/or 18; bivalent\* types represent 16, 18, 31, 33 and 45, and nonavalent: 6, 11, 16, 18, 31, 33, 45, 52 and 58. Types included in each of the groups is described in the methods part. Due to the nature of the analysis, infections could be counted more than once as some samples were infected by multiple HPV types.

Of the 185 cases of anal cancer, 164 (88.65%) samples were positive for at least one HPV type. High-risk types were detected in 87.03% of the samples. Mono-infection of HPV 16 was present in 145 (83.2%) samples. Other hr-HPV types were detected in 4.3% of the samples while low-risk HPV were present in 1.6%. HPV 18 was the second most dominant type, present in 3 samples (1.62%). Presence of low-risk types without any other hr-HPV was detected in 3 samples (1.62%) (Figure 25).



**Figure 25. HPV positivity in 185 anal cancer samples collected between 2009 to 2018.** High-risk types included: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68, other-hr-HPV types (35, 39, 51, 56, 59 and 68) and low-risk HPV types 6, 11, 40, 42, 43, 54, 44 and 61). Due to the nature of the analysis, infections could be counted more than once as some samples were infected by multiple HPV types.

#### 4.3.4 Vaccine preventable fraction

For the anal cancer cohort, bivalent vaccine could potentially prevent 157/185 (84.86%). When cross-reactivity is considered, 158/185 (85.40%) are potentially preventable. The quadrivalent and nonavalent vaccine preventable fraction for AC was 160/185 (86.49%) and 162/185 (87.57%) respectively. If consider together AIN + AC, the bivalent vaccine could potentially prevent 84.5% of anal lesions, 85.0% if cross-reactivity considered, 86.5% by the quadrivalent, and 88% by the nonavalent vaccines.

#### *4.3.5 Is there an association with HPV positivity and demographic characteristics/clinical factors in anal cancer?*

Overall HPV status by demographics and clinical cohorts are presented in Table 18. Anal cancer was more prevalent in females (64.86%) than males (35.13%). Majority of cases were diagnosed in individuals 60-69 (30.27%) and majority of cases were stage II and III (36.75% and 28.64%). Additionally, 74.59% of cases responded to treatment and 67.02% were alive at date of censoring.

Of the female cases 90.83% were HPV positive; of the male cases, 84.61% were HPV positive. The majority of HPV positive cases were diagnosed in the 60-69 (31.09%) age range and at stage II (37.80%), 76.21% responded to treatment and 70.12% were alive at data censoring. No significant changes in HPV positivity were identified during in the 10-years period analysed ( $p=0.100$ ).



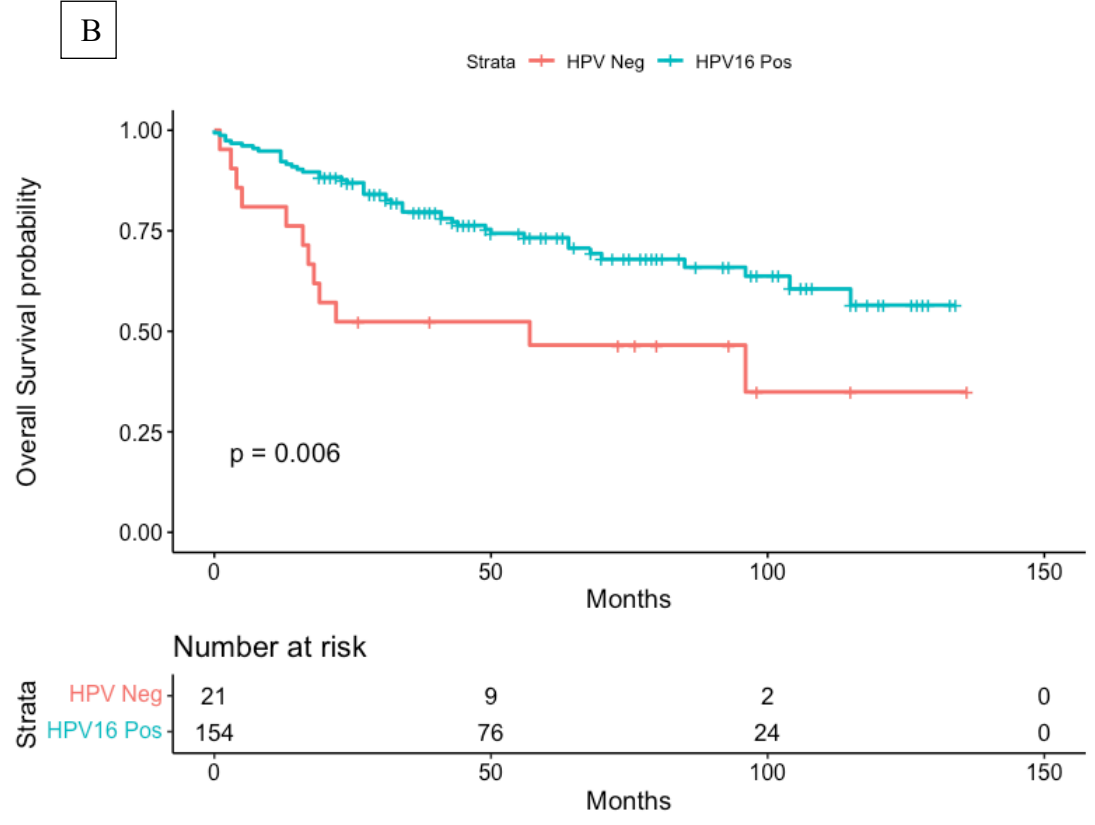
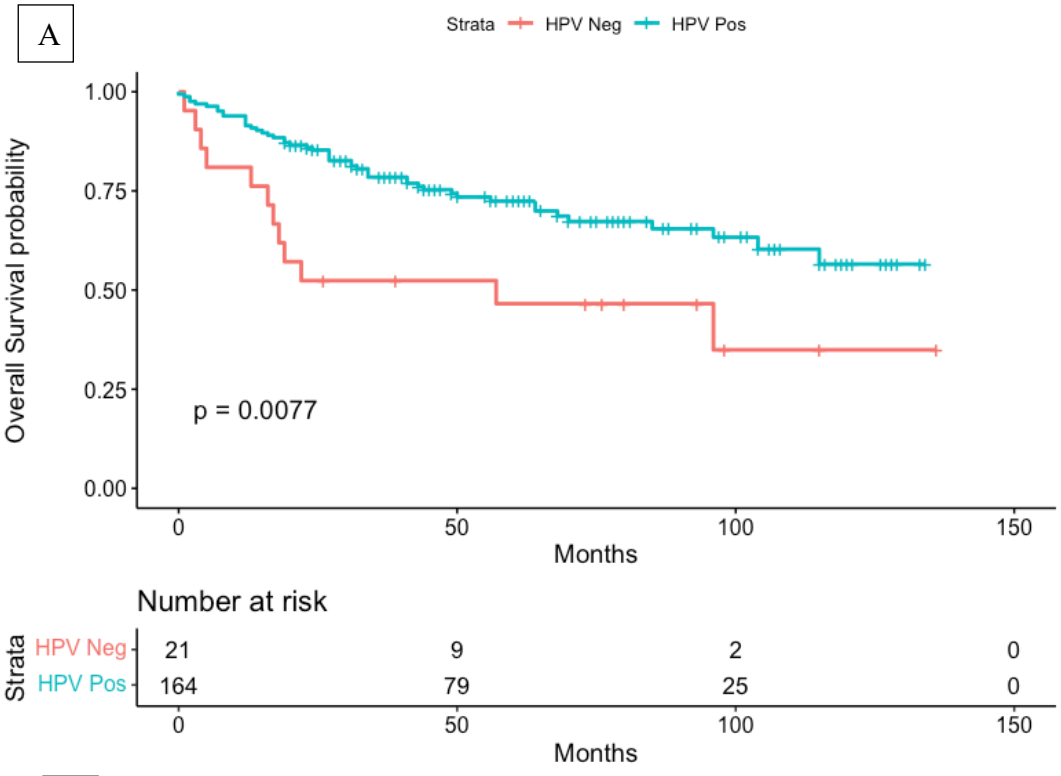
**Table 18. Demographic & clinical characteristics of the anal cancer samples collected between 2009 to 2018 in the South-East of Scotland.** Cases are also stratified according to whether they were HPV positive or negative or according to their HPV 16 virus load. “N” corresponds to the total number of samples for every category described above.

		HPV status & survival samples cohort n(%)			Viral load & survival samples cohort n (%)			
Variable	Level	n = 185	HPV-positive (n=164)	HPV-negative (n=21)	n = 145	Low (n=47)	Medium (n=50)	High (n=48)
<b>Sex</b>	<b>Female</b>	120 (64.86%)	109 (66.46%)	11 (52.38%)	101 (69.65%)	29 (61.70%)	38 (76.0%)	34 (70.83%)
	<b>Male</b>	65 (35.13%)	55 (33.53%)	10 (47.61%)	44 (30.34%)	18 (38.29%)	12 (24.0%)	14 (29.16%)
<b>Age</b>	<b>&lt;50</b>	28 (15.13%)	25 (15.24%)	3 (14.28%)	22 (15.17%)	7 (14.89%)	6 (12.0%)	9 (18.75%)
	<b>50 - 59</b>	48 (25.94%)	48 (29.26%)	0 (0%)	44 (30.34%)	18 (38.29%)	14 (28.0%)	12 (25%)
	<b>60 - 69</b>	56 (30.27%)	51 (31.09%)	5 (23.80%)	45 (31.03%)	17 (36.17%)	12 (24.0%)	16 (33.33%)
	<b>70 and over</b>	53 (28.64%)	40 (24.39%)	13 (61.90%)	34 (23.44%)	5 (10.63%)	18 (36.0%)	11 (22.91%)
<b>Collection year</b>	<b>2009</b>	14 (7.567%)	13 (7.93%)	1 (4.761%)	12 (8.28%)	4 (8.51%)	3 (6.00%)	5 (10.41%)
	<b>2010</b>	12 (6.49%)	10 (6.10%)	2 (9.523%)	9 (6.21%)	2 (4.25%)	6 (12.00%)	1 (2.08%)
	<b>2011</b>	15 (8.11%)	13 (7.93%)	2 (9.523%)	13 (8.96%)	6 (12.76%)	2 (4.00%)	5 (10.41%)
	<b>2012</b>	17 (9.19%)	13 (7.93%)	4 (19.04%)	11 (7.59%)	3 (6.38%)	3 (6.00%)	5 (10.41%)
	<b>2013</b>	16 (8.65%)	14 (8.54%)	2 (9.523%)	10 (6.90%)	3 (6.38%)	4 (8.00%)	3 (6.25%)
	<b>2014</b>	18 (9.73%)	15 (9.15%)	3 (14.28%)	15 (10.34%)	7 (14.89%)	6 (12.00%)	2 (4.17%)
	<b>2015</b>	25 (13.51%)	22 (13.41%)	3 (14.28%)	18 (12.41%)	7 (14.89%)	7 (14.00%)	4 (8.33%)
	<b>2016</b>	22 (11.89%)	22 (13.41%)	0 (0.00%)	20 (13.79%)	3 (6.38%)	6 (12.00%)	11 (22.91%)
	<b>2017</b>	23 (12.43%)	20 (12.19%)	3 (14.28%)	17 (11.72%)	6 (12.76%)	5 (10.00%)	6 (12.5%)
<b>2018</b>	23 (12.43%)	22 (13.41%)	1 (4.761%)	20 (13.79%)	6 (12.76%)	8 (16.00%)	6 (12.5%)	
<b>Stage</b>	<b>I</b>	27 (14.59%)	23 (14.02%)	4 (19.04%)	22 (15.17%)	5 (10.63%)	8 (16.00%)	9 (18.75%)
	<b>II</b>	68 (36.75%)	62 (37.80%)	6 (28.57%)	54 (37.24%)	15 (31.91%)	17 (34.00%)	22 (45.83%)
	<b>III</b>	53 (28.64%)	48 (29.26%)	5 (23.80%)	40 (27.58%)	15 (31.91%)	16 (32.00%)	11 (22.91%)
	<b>IV</b>	35 (18.91%)	23 (14.02%)	6 (28.57%)	25 (17.24%)	11 (23.40%)	9 (18.0%)	5 (10.41%)
	<b>Unknown</b>	2 (1.08%)	2 (1.22%)	0 (0.00%)	2 (1.37%)	1 (2.12%)	0 (0.00%)	1 (2.08%)
<b>Response to Treatment</b>	<b>Yes</b>	138 (74.59%)	125 (76.21%)	13 (61.90%)	111 (76.55%)	35 (74.46%)	37 (74.00%)	39 (81.25%)
	<b>No</b>	34 (18.37%)	28 (17.07%)	6 (28.57%)	23 (15.86%)	8 (17.02%)	9 (18.00%)	6 (12.50%)
	<b>Unknown</b>	13 (7.03%)	11 (6.71%)	2 (9.52%)	11 (7.59%)	4 (8.51%)	4 (8.00%)	3 (6.25%)
<b>Vital status</b>	<b>Alive</b>	124 (67.02%)	115 (70.12%)	9 (42.85%)	104 (71.72%)	29 (61.70%)	33 (66.00%)	42 (87.5%)
	<b>Deceased</b>	61 (32.97%)	49 (29.87%)	12 (57.14%)	41 (28.27%)	18 (38.29%)	17 (34.00%)	6 (12.50%)

#### 4.3.6 Does HPV positivity have an impact on survival in patients with anal cancer?

Of the 185 anal cancer cases included in the qualitative analysis, 61 (32.97%) patients had died during follow up. Eighty-point three percent of the deceased patients were HPV positive while 92.74% of those still alive were positive for HPV. Kaplan-Meier curves were generated, stratifying by HPV status (positive and negative). (Figure 26, A), and HPV 16 status (Figure 26, B), which shows HPV positive status and HPV 16 positive status were linked to improved survival (log-rank test p value 0.0077 and 0.006 respectively).

The univariate analysis showed that HPV positive status was associated with a better overall survival, with a hazard ratio (HR) of 0.44 (95% CI 0.23 – 0.82,  $p=0.01$ ) (Table 19) . In the univariate Cox model, variables associated with worse overall survival were Stage III; HR 5.0 (95% CI 1.1 – 22),  $p=0.003$  and Stage IV HR 25.2 (5.65 – 109),  $p<0.001$  vs. stage I and no response to treatment 0.12 (95% CI 0.07 - 0.21)  $p<0.001$  vs. response to treatment. After adjusting for age, gender, stage and response to treatment, HPV status continued to influence the overall survival, HR 0.24 (95% CI 0.11 – 0.55)  $p<0.001$ . When adjusting for age and gender alone, HR for HPV positive status was 0.41(95% CI 0.21 – 0.82)  $p=0.011$  (Table 20).



**Figure 26. Overall Survival probability for “any” HPV positive and HPV negative (A) and for HPV 16-positive vs. HPV-negative (B) anal cancer cases using Kaplan Meier estimator.** Survival time expressed in days from diagnosis date. Data censored in July 2020.

**Table 19. Univariate and multivariate hazard ratio of HPV status derived using Cox regression for anal cancer cases.**

Variable	Level	Unadjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value
HPV	HPV Neg	1		1		1	
	HPV Pos	0.44 (0.23 - 0.82)	0.01	0.24 (0.11 - 0.55)	<0.001	0.41 (0.21 - 0.82)	0.011
Sex	Male	1		1		1	
	Female	0.85 (0.51 - 1.4)	0.549	0.98 (0.52 - 1.87)	0.955	0.90 (0.53 - 1.53)	0.704
Age	<50	1				1	
	50 - 59	1.6 (0.68 - 3.7)	0.288	1.09 (0.43 - 2.79)	0.852	1.84 (0.77 - 4.37)	0.167
	60 - 69	1.2 (0.53 - 2.9)	0.635	2.40 (0.97 - 5.98)	0.059	1.31 (0.56 - 3.06)	0.532
	70 and over	1.6 (0.70 - 3.7)	0.257	1.88 (0.69 - 5.11)	0.217	1.48 (0.63 - 3.45)	0.365
Stage	I	1		1			
	II	3.5 (0.8 - 15)	0.095	4.28 (0.96 - 19.13)	0.057		
	III	5.0 (1.1 - 22)	0.003	5.67 (1.24 - 25.93)	0.025		
	IV	25.6 (6.0 - 109)	<0.001	18.58 (3.96 - 87.18)	<0.001		
Response to treatment	No	1		1			
	Yes	0.12 (0.07 - 0.21)	<0.001	0.16 (0.07 - 0.33)	<0.001		

#### 4.3.7 HPV 16 viral load in the anal cancer cohort.

As described before, viral load in cancer samples was assessed by ddPCR calculating the number of copies of the L1 gene in those with HPV 16 positive anal cancers. A total of 145/154 samples (94.15%) were associated with valid reads in the ddPCR for HPV 16 L1 sequences. Nine samples were excluded from the analysis because they generated less than 10,000 droplets even after repeat testing.

The individual HPV 16 viral loads were ranked from smallest to largest and separated using tertiles. The range of viral loads observed was 0.021 to 710 copies of the HPV L1 gene per cell, with an average of 60.57 L1 copies. Of those alive at the time of data censoring, 27.88% (20.17 – 37.17) cancer samples were associated with a low viral load, 31.73% (23.57 – 41.19) a medium viral load and 40.38% (31.46 – 49.99) a high viral load. Comparatively, in those who died low viral load was present in 43.9% (29.89 – 58.96), medium viral load in 41.46% (27.75 – 56.63) whereas 14.63% (6.88 – 28.44) had a high number of copies. Viral load values are described in Table 20.

**Table 20. Level of viral loads by vital status obtained in the HPV 16-positive group anal cancer cases.**

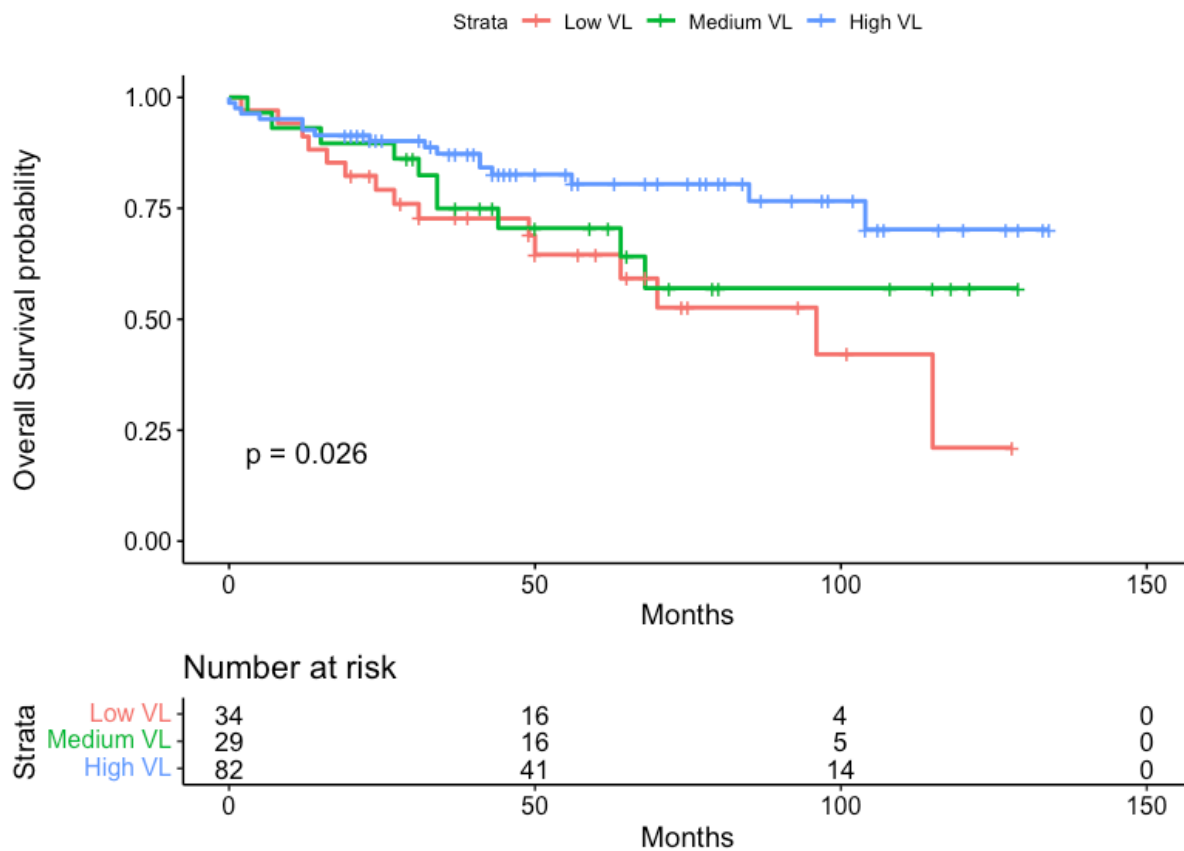
	<b>Alive</b>	<b>Deceased</b>
<b>Low VL</b>	29 (27.88%)	18 (43.9%)
<b>Medium VL</b>	33 (31.73%)	17 (41.46%)
<b>High VL</b>	42 (40.38%)	6 (14.63%)
<b>Total</b>	104	41

	<b>Alive</b>	<b>Deceased</b>
<b>Low VL</b>	29 (27.88%)	18 (43.9%)
<b>Medium VL</b>	33 (31.73%)	17 (41.46%)
<b>High VL</b>	42 (40.38%)	6 (14.63%)
<b>Total</b>	104	41

Those who were deceased had a median L1 viral load of 33.11 (IQR 5.87 - 83); while those still alive had a median number of copies of 74.09 (IQR 5.80 – 78.19). A total of 52 samples (33.99%) had a low viral load ( $\leq 5.57$ ), 52 medium (33.99%) (5.58 – 25.63) and 49 (32.03%) high viral load ( $> 25.64$ ) copies of L1 gene per samples.

#### 4.3.8 Survival analysis

For the Kaplan-Meier estimator, overall survival was calculated by viral load groups: low, medium and high viral load. The overall survival rate was found to be higher in patients with medium and high viral loads as compared to those with low viral loads, with a p-value of 0.026 (Figure 27).



**Figure 27. Kaplan-Meier survival curve stratified by viral load (Low and Medium/High).** Survival time expressed in months from the diagnosis date. Data censored on 31<sup>st</sup> July 2020.

Table 21 displays the overall survival rate categorized by the demographic and clinical variables outlined in Table 17, with the viral load divided into three tertiles and the low viral load group serving as the reference. The univariate analysis demonstrated that high viral load was associated with a better overall survival rate, with a hazard ratio of 0.28 (0.11 – 0.71, p=0.007). The variables that were linked to a worse

overall survival rate were stage IV cancer (compared to stage I), with a hazard ratio of 25.2 (95% CI 5.65 – 113) and a p-value of less than 0.001, as well as a lack of response to treatment, with a hazard ratio of 0.13 (95% CI 0.064 - 0.27) and a p-value of less than 0.001.

However, when Cox HR was adjusted, viral load did not significantly influence the overall survival.; medium viral load HR 1.04 (0.45 – 2.40) p=0.924, high viral load 0.39 (0.12 – 1.24) p=0.111. Only stage IV HR 21.52 (4.01 - 115.41) p<0.001 and response to treatment HR 0.23 (0.09 - 0.56) p=0.001 were associated with a worse survival in the adjusted analysis. When stage and response to treatment were not included in the adjustment, high viral load had an impact in the overall survival (0.27, 0.11 – 0.68) p=0.006.

**Table 21. Overall survival stratified by clinical variables, demographic variables and L1 viral load.** HR derived using Cox regression. Adjusted for sex, age, stage, and response to treatment and with only demographics<sup>170</sup>.

Variable	Level	Unadjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value
Viral Load	Low	1		1		1	
	Medium	0.91 (0.47 – 1.76)	0.774	1.04 (0.45 – 2.40)	0.924	0.80 (0.40 – 1.60)	0.531
	High	0.28 (0.11 – 0.71)	0.007	0.39 (0.12 – 1.24)	0.111	0.27 (0.11 – 0.68)	0.006
Sex	Male	1		1		1	
	Female	1.2 (0.6 - 2.4)	0.625	1.09 (0.47 - 2.53)	0.838	1.35 (0.65 – 2.76)	0.419
Age	<50	1		1		1	
	50 - 59	1.51 (0.58 - 4.0)	0.398	0.77 (0.25 – 2.32)	0.639	1.30 (0.50 - 3.41)	0.588
	60 - 69	0.94 (0.34 - 2.6)	0.912	2.20 (0.69 - 7.01)	0.183	0.85 (0.30 - 2.40)	0.753
	70 and over	1.53 (0.56 - 4.2)	0.405	3.05 (0.757 - 12.31)	0.117	1.65 (0.58 - 4.65)	0.347
Stage	I	1		1			
	II	2.2 (0.48 - 10)	0.302	2.31 (0.48 – 11.18)	0.299		
	III	2.9 (0.62 - 14)	0.178	2.58 (0.50 - 13.23)	0.254		
	IV	25.2 (5.65 - 113)	<0.001	21.52 (4.01 - 115.41)	<0.001		

Response to treatment	No	1		1	
	Yes	0.13 (0.064 - 0.27)	<0.001	0.23 (0.09 - 0.56)	0.001

#### 4.4 Discussion

This chapter has looked at the HPV prevalence in a population-based cohort of anal cancer collected during a 10-year period. Additionally, HPV status and load and its influence on overall survival in anal cancer samples was assessed.

The majority of anal cancer samples were positive for at least one HPV type (88.65%) HPV 16 was the clear dominant type (83.24%) in the anal cancer positive cases. This finding is consistent with the high rate of HPV positivity observed in anal cancer, as well as the high prevalence of HPV 16, as reported by Desanjose *et al.*<sup>131</sup>.

Data shows most of anal lesions could be potentially preventable by the HPV vaccines offered in the UK in 2023. Anal cancer data shows that, given the massive dominance of HPV 16 and 18, the incremental benefits of the different vaccine formulations are minimal. Still, it will take some time to see the full effectiveness of the vaccine and the reduction in lesions and cancers due to the age where majority of lesions have identified. Majority of anal cancer cases have been diagnosed in >50 years old<sup>84,171</sup> (84.86%), almost 60% in >60-year-olds. Given the time frame of the vaccination (2008 for schoolgirls, 2019 for schoolboys and 2017 for MSM up to 45 years old) and the peak prevalence of the anal disease, it would take at least two decades to see the full effect of the HPV immunisation in the population.

HPV positive prevalence in women was higher than in males (66.46% vs 33.53%). This is consistent with published data, including that the Cancer Research UK and Moscicki *et al.*<sup>171,172</sup>. Furthermore, the study



found that HPV status (HPV-positive) was linked to a better overall survival rate. Hazard ratio was 0.24 and a p-value of 0.001, when compared to cases where HPV status was negative. This find is consistent with the systematic review conducted by Urbute *et al.*, which showed that HPV-positive anal cancer cases had significantly better overall survival compared with HPV-negative<sup>161</sup>. This is similar to an emerging pattern in other cancers associated with HPV, including cervical<sup>173,174</sup>, oropharyngeal<sup>175,176</sup>, penile<sup>177,178</sup> and vulval cancers<sup>179</sup>.

The ddPCR assay showed that a high viral load, as measured by quantifying HPV 16 L1 gene copies, was associated with better clinical outcomes than low copies of L1 in the univariate analysis (HR 0.28, 95% CI 0.11 – 0.71, p=0.007) when compared with low and medium viral load. However, when the Cox hazard ratio was adjusted, viral load did not appear to influence overall survival (HR 0.39, 95% CI 0.12 – 1.24, p=0.111). One of the reasons behind this, could be due to the relatively small sample size in the study. As the confidence interval just exceeds 1, it would be plausible that with a larger study, it may tip into significance. When clinical variables are not considered in both the unadjusted and adjusted analysis, viral load, and survival data, where high viral load correlated with better survival. Moreover, viral load association with survival in anal cancer ties with other HPV driven cancers, including cervix<sup>148,149</sup> and oropharynx<sup>150,151,153,155,180</sup>. Association of high viral load of HPV in anal cancer and better survival indicates that could be used as a biomarker for better anal cancer optimal management strategies.

The complete understanding of why high viral load in cancers may be associated with a better prognosis is yet to be achieved. It is possible that cancer samples with a lower HPV viral load may have other factors responsible for the cancer, such as impairment of viral function through integration and/or epigenetic mechanisms<sup>72</sup>, thus, the viral genome status could significantly impact patient prognosis. Studies have shown that head and neck cancer cases with an episomal status of HPV have a more favourable prognosis

<sup>180-183</sup>, and that episomal status is associated with high viral load, while integration of the HPV genome is associated with low viral load <sup>180-183</sup>. This suggests that the detection of the L1 target alone may not provide comprehensive results due to a possible reduction of L1 presence in the HPV genome resulting from integration events. The use of a combination of targets (L1 + E6 or E7) could improve the detection of viral load in cases where the HPV genome has integrated into the host.

There are limitations to the study. Although the sample set was well annotated it was still relatively small and the AIN sample size was too small for any formal statistical analysis. Additionally, due to the significant relationship of HIV and immune alterations with anal cancer, knowledge of HIV and immune status could have provided more contextualisation of the viral load and anal cancer. It would also be very interesting to look at the influence of viral load on lesion progression in a prospective study and determine if viral load can be used as a biomarker that we could use to distinguish between lesions that would lead into neoplasms and those that would eventually be cleared.

Nevertheless, given the results of HPV status and overall survival, it could be argued that assessment of HPV status in anal cancer cases and viral load is worthy to be considered for anal lesions management in Scotland. Similar High HPV prevalence detected in cervical cancer and the reduction already registered in the number of lesions due to the vaccine, suggest that the majority of anal lesions could also be prevented by the HPV vaccines and in the future. However, as mentioned above, it will take one or two decades to see the full impact.

This chapter has provided with an insight into the proportion of HPV anal cancers that are associated with HPV in anal cancers from Scotland, and the HPV dominant types. In addition, HPV 16 viral load analysis suggest that high viral load could be associated with better survival, however further studies to investigate

this would be of value including studies where additional targets to L1 are used. As described in the systematic review by Theophanous *et al.* (2022) of biomarkers for anal cancer treatment, there are only a few prognostic factors that can help predicting anal cancer outcomes, including the presence of leukocytosis, neutrophilia, and anemia at baseline measurement<sup>184</sup>.

In addition to ddPCR, next generation sequencing has also started to be used as an HPV diagnostic tool in more laboratories. The deep analysis of HPV genome could result in the identification of other small variabilities/mutations in the HPV genome that could potentially be used as biomarkers. Over the last 5-8 years, HPV sub-lineage has been identified as a potential biomarker for cancer risk.

It is clear the necessity for biomarkers of significant disease for anal lesions. Due to the lack of data available for anal cancers, in the subsequent chapter it will be explored the whole genome of type HPV 16 in anal cancers and determine if sub-lineage can be used as a biomarker.

## 5. Optimisation of molecular and bioinformatic tools to support the identification of HPV 16 sub-lineages

### 5.1 Introduction

Thanks to the data collected in Chapter 4, we now know that HPV was present in a large proportion of anal cancer cases collected in the East of Scotland between 2009 and 2018. Using a PCR-based assay, HPV was detected in 89% of anal cancers. The dominant type identified was HPV 16, the higher carcinogenic risk of all the hr-HPV types due to the activity of the oncoproteins E6 and E7<sup>185</sup>.

Different investigators have assessed in depth the HPV 16 type genome in cervical cancer. For example, Mirabello *et al.* identified that some HPV 16 sub-lineages are associated with a higher risk of high-grade lesions or cervical cancer<sup>31</sup>. Other studies have also looked at the SNP of HPV 16, identifying that SNP 350 is associated with a higher persistence of infection<sup>51</sup>. But most available studies have been performed on cervical lesions and very little has been done on anal samples. Therefore, part of this PhD thesis aimed to investigate the variability of the HPV 16 genome (lineages and sub-lineages identification) present in anal cancer samples.

However, before starting with the sequencing, it was necessary to identify the best protocols for nucleic acid extraction from the anal cancer samples, next-generation sequencing (NGS) library preparation and bioinformatic analysis.

### *5.1.1 Formalin fixed paraffin embedded (FFPE) nucleic acid extraction for Next Generation Sequencing downstream.*

As described in the previous chapter (Chapter 4), the available specimen for any application of NGS to genomically characterise HPV associated with anal cancer samples was anal lesions preserved as formalin-fixed paraffin-embedded (FFPE). In pathology labs, FFPE specimens are still one of the most abundant sources of clinical material. However, the fixation process is known to degrade nucleic acid during the process and therefore compromise and make more difficult the subsequent analysis<sup>186</sup>. Formaldehyde induces oxidation and deamination reactions and the formation of cross-links in the nucleic acid. These chemical modifications can hinder sequencing by inhibiting the molecular reactions during the library prep step. They may incur direct changes at a single genome base (high number of Ns) or lead to DNA fragmentation that renders sequencing and analysis challenging. Moreover, it is also known that the FFPE nucleic acid quality/adequacy can vary widely due to fixation conditions and fixation process, the nature and extent of which may have a variable impact on downstream NGS analyses<sup>187</sup>. However, recent developments in nucleic acid extraction methods for FFPE include new deparaffinisation solutions, DNA sequence repair strategies and magnetic bead technology that help obtain the most from fixed samples.

Most HPV genotyping assays in the market are not validated for FFPE material. However, different publications have evaluated their feasibility for FFPE<sup>188–191</sup>. These tests typically target the L1 or E6/E7 genes and have a PCR step before detection. The target size of these assays is usually small, which could explain why they are not affected by the DNA fragmentation and cross-links generated during the FFPE preservation. On the other hand, whole genome sequence analysis by NGS could be affected by excessive fragmentation or low quality of the nucleic acid material.

Multiple publications have analysed nucleic extraction methods and how these affect molecular HPV detection using PCR<sup>85,191–195</sup>. However, there is a lack of information on the comparative performance of FFPE extraction technologies to support the detection of HPV for molecular applications, including NGS.

### *5.1.2 NGS approach selection*

There are two different approaches for NGS, short read and long read. Short-read NGS can use fragments between 150 – 1500 bp, while long-read sequencing uses DNA lengths of several kb<sup>196</sup>. The selection of the NGS approach may be influenced by the type of specimen available. Some specimens could have a more fragmented DNA, making them unsuitable for long-read sequencing. It is again the case of FFPE, where the preservation process also impacts the length of DNA by fragmenting the DNA and reducing the amplifiable proportion<sup>197</sup>.

Also, most clinical diagnostic laboratories currently do not culture HPV from samples. HPV detection is performed directly from the specimen preserved. As samples contain organisms other than HPV (including human cells, bacteria and other viruses), it is impossible to obtain a pure isolate of the HPV present in the sample. Thus, specimen/sample type and a low ratio viral:human genome in the sample. As the aim is to identify HPV 16 sub-lineages, the whole genome of the virus must be sequenced. To achieve this, a target enrichment protocol that could amplify the entire length of the HPV 16 genome was selected. This enriched protocol increases the probability of generating the necessary reads and the highest coverage to obtain a complete illustration of the entire HPV 16 genome, leading to lineage/sub-lineage, integration, or mutation analysis.

### 5.1.3 Bioinformatic Tools for HPV Analysis

Bioinformatics tools have been designed to manipulate large amounts of data in the easiest way possible. Many tools are available for multiple types of analysis, most publicly available. For HPV, some tools are also available designed for HPV identification from raw sequencing files (FASTQ files)<sup>198,199</sup>, but these tools have been developed as student projects and lack consistent technical support. In terms of an agreed pipeline by the HPV community, at time of preparation of this thesis the author is not aware of any guidelines or documents that stipulate how to analyse HPV through bioinformatics tools or quality parameters. Therefore, a new bioinformatic analysis protocol was required.

#### 5.1.4 Aim

The chapters aim to optimise the nucleic acid extraction method, prepare an NGS protocol for whole genome sequencing of HPV 16 and set up and validate NGS bioinformatic pipeline for identifying HPV 16 sub-lineages in anal FFPE samples.

- Determine the optimal extraction method of FFPE for molecular detection of HPV 16.
- Prepare a protocol to perform whole genome sequencing of HPV 16.
- Perform validation of the NGS/bioinformatic protocol through inter-laboratory comparison.

## 5.2 Material and Methods:

Nucleic acid extraction comparison for whole genome sequencing

### 5.2.1 Extraction comparison

Forty-eight anal samples used in Chapter 4 were used to extract nucleic acid using three different methods. The extraction methods used included:

- Manual extraction method based on Qiagen DNA Mini Kit.
- GeneRead DNA FFPE Kit.
- Automated extraction method using the Seegene Nimbus platform.

The process and characteristics for each of the methods is described in the general materials and methods (Chapter 2)

### *5.2.2 DNA quantification: Qubit and qPCR*

Sample DNA obtained from the three different extraction methods were quantified using the Qubit and an HPV 16 qPCR designed specially to have amplicons similar to the size of DNA fragments valid for Illumina sequencing. Qubit and qPCR methods are described in the general methods chapter (Chapter 2).

### *5.2.3 analysis*

DNA quantity and PCR Ct values were assessed for samples prepared using the difference three extraction methods. To determine if statistically significant differences exist between extraction kits, analysis of variance (ANOVA) and Tukey HSD statistics were calculated. These statistics and the boxplots were calculated using R, version 1.3.1093

### *5.2.4 HPV 16 sequencing method of choice*

The NGS method was selected considering the instruments available in the Royal Infirmary of Edinburgh laboratories, specimen type, protocols available and cost. For the bioinformatic part, analysis requirements and infrastructure accessible for the WGS analysis was assessed.



Initially, libraries were prepared using the Illumina TruSeq DNA nano kit. However, due to low DNA yield, the TruSeq kit was replaced by the DNA prep kit (Illumina).

*5.2.5 validation of the sequencing and bioinformatic – Agreement analysis Karolinska Institute*

A comparison of the identified sub-lineages was required to validate the sequencing protocol and bioinformatics tools prepared. A total of 25 anonymised fastq files were sent to the Karolinska Institute (Dr Sara Arroyo-Mühr) for a blind comparison. No information other than the samples' HPV 16 status was shared. The twenty-five samples selected consisted of twenty-one identified as HPV 16 A1, three as A2, and one as A3.

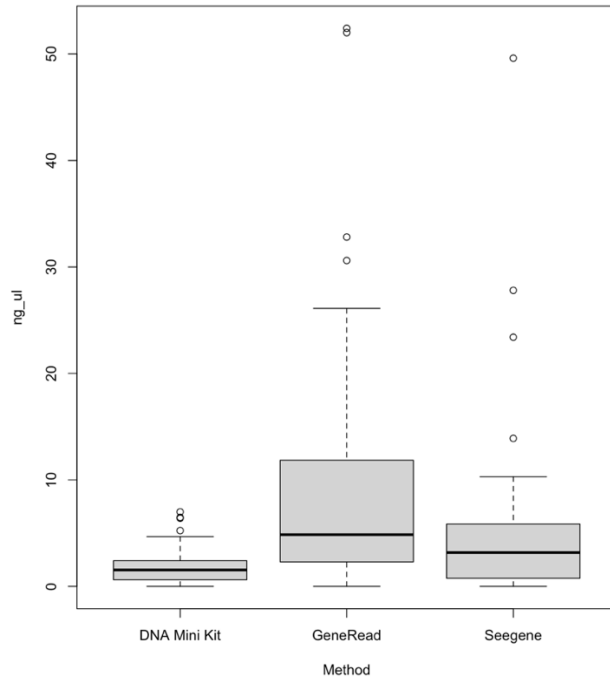
**5. 3 Results:**

*5.3.1 DNA concentration*

From the total 48 samples, four samples extracted using the GeneRead did not yield detectable DNA by the Qubit, six by the DNA Mini Kit and seven by the Seegene Universal. The average DNA quantity detected in GeneRead was 10.64 ng/μl, 2.14 ng/μl from DNA mini kit and 6.16 ng/μl from the Seegene method. The range was 0.74 – 52.4 for the GeneRead, 0.42 – 7 for the DNA Mini kit and 0.4 – 49.6 for the Seegene extraction (Table 22). Figure 28 represents the median and outliers for DNA concentration obtained from each extraction method.

**Table 22. Minimum, maximum and average values were obtained from Qubit extraction and qPCR for the three different extraction methods.**

	Qubit ng/μl		qPCR Ct	
	Max - Min	Average	Max - Min	Average
<b>GeneRead</b>	0.74 - 52.4	10.64	17.36 - 33.68	23.84
<b>DNA Mini Kit</b>	0.42 - 7	2.14	19.93 - 33.92	24.67
<b>Seegene</b>	0.4 - 49.6	6.16	18.11 - 32.41	23.92



**Figure 28. Boxplot representing the quantity of DNA (ng/ml) obtained (n=48) from the three different extraction methods**

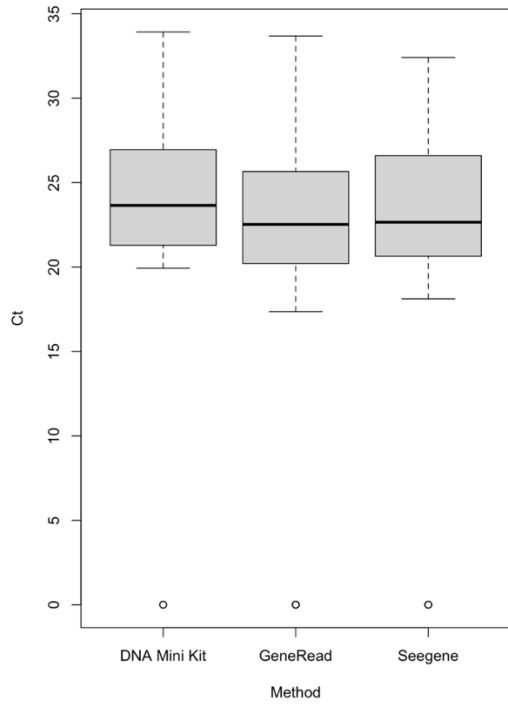
ANOVA found significant differences between the three assays ( $p < 0.001$ ). Tukey HSD test found significant differences between the GeneRead and DNA mini kit ( $p < 0.001$ ) and the GeneRead and Seegene (0.031). However, no significant differences were identified between the Seegene and the DNA Mini Kit ( $p = 0.136$ ).

### 5.3.2 qPCR

From the total 48 samples, GeneRead extracted samples did obtain a valid (logistic curve between cycles 10 – 40) cycle threshold (Ct) value in 44 samples, with four not providing any amplification. Samples extracted by the DNA Mini Kit and Seegene Nimbus had 46 and 45 valid samples, respectively. The average Ct values obtained from the three methods were 23.84, 24.67 and 23.92 for GeneRead, DNA Mini Kit and Seegene, respectively. Range Ct values for GeneRead were 17.36 - 33.68, 19.93 - 33.92 and 18.11 - 32.41 for GeneRead, DNA Mini kit and Seegene, respectively. Values are described in Table 24. All Ct values

obtained for the 48 samples are described as a boxplot in Figure 29, representing the median and outliers of Ct values obtained from each extraction method.

Statistical analysis could not find significant differences in the qPCR Ct value between the extraction methods ( $p=0.428$ ). Tukey HSD test could not detect differences between any of the methods: GeneRead vs DNA mini kit ( $p=0.409$ ), DNA Mini Kit vs Seegene ( $p=0.660$ ) and GeneRead vs Seegene ( $p=0.911$ ).



**Figure 29. Boxplot representing the Ct value obtained (n=48) from the three extraction methods.**

### 5.3.3 NGS protocol for WGS of HPV 16

A PCR-based enrichment approach was selected to amplify the HPV present in the samples and reduce the increase in the ratio of HPV:human nucleic acid present in the sample. To amplify the entire genome of the HPV, a 47-amplicon approach was selected. A large number of primers and a small size of the amplicons (~200 bp size) were chosen to reduce the chances of mismatch due to the fragmentation of the FFPE samples.

Due to the availability of an Illumina MiSeq (San Diego, USA) in the reference laboratories and the potential fragmented status of the sample DNA, a short-read sequencing method was selected. The average fragment size obtained with the Illumina DNA prep kit was 250-300 bp in the different runs performed.

The infrastructure available at NHS Lothian and the requirements to perform the desired analysis were considered for the bioinformatic analysis. Due to the absence of accessibility to an analysis cluster, WGS analysis needed to be performed on a local machine.

As there were no already designed pipelines for HPV sub-lineage identification, it was put together the bioinformatic tools necessary for the analysis. This included the mapping of the sequencing reads against an HPV 16 reference, read and coverage analysis, consensus sequence obtaining, alignment, and phylogenetic analysis. Chapter 2 includes the full description of the tools used.

### 5.3.4 Selection of library kit

The library prep kit of choice was the TruSeq DNA nano kit (Illumina), as used in the Arroyo *et al.* publication<sup>200</sup>. Initially, libraries were prepared from anal cancers using the Illumina TruSeq DNA nano. However, libraries did not contain any DNA, or the DNA concentration ng/ $\mu$ l was very low (<0.8), meaning it was not possible to reach the recommended pool concentration for Miseq analysis (4  $\mu$ M). Figure 30 to

Figure 32 represent different DNA concentrations and size examples for the libraries prepared with the TruSeq DNA nano kit.

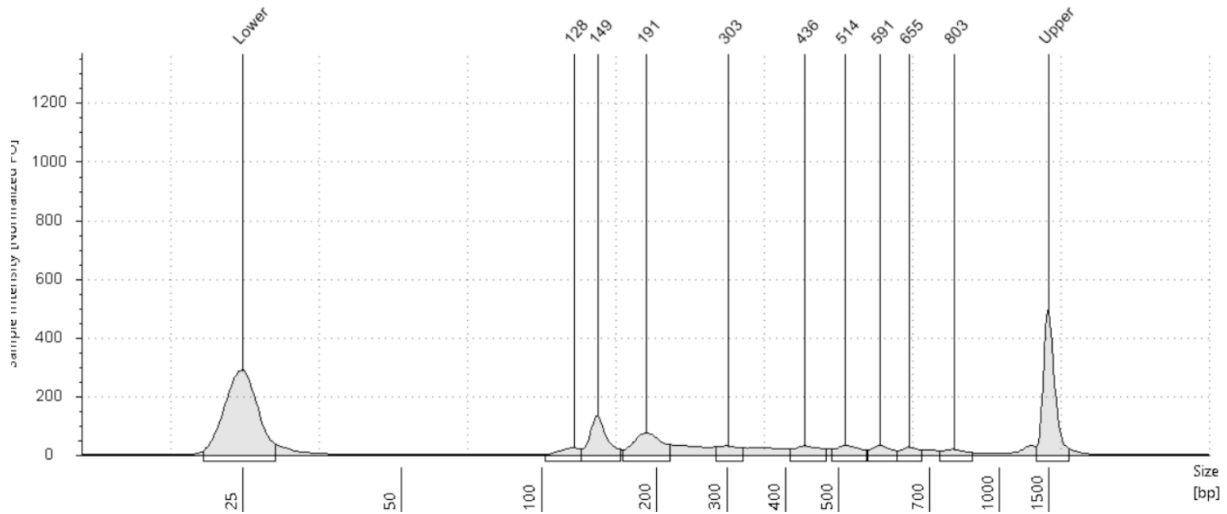


Figure 30. Tape-station graph representing the sample intensity (FUI) for the DNA library's different sizes (bp). This case shows a sample with poor concentration.

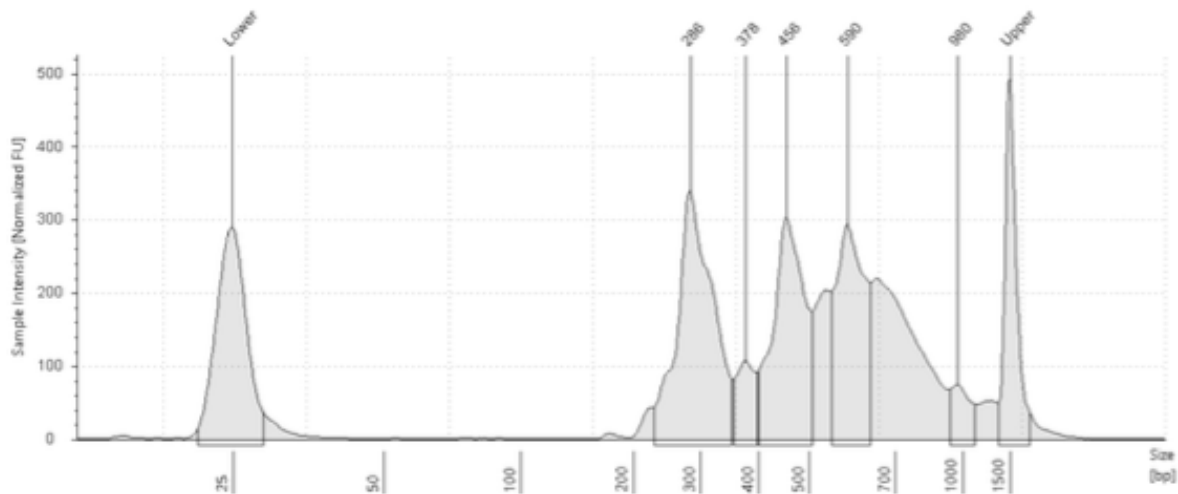
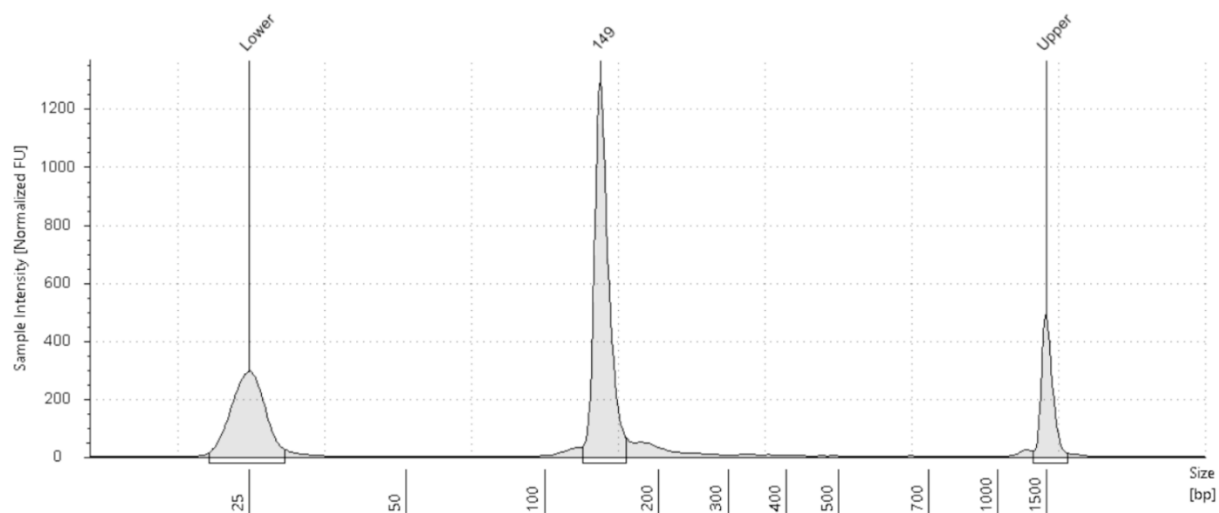


Figure 31. Tape-station graph representing the sample intensity (FUI) for the DNA library's different sizes (bp). This case shows a sample with good intensity. However, target size was not achieved as other DNA sizes are present in the sample.



**Figure 32. Tapestation graph representing the sample intensity (FU) for the DNA library's different sizes (bp).** This case shows a sample with no target DNA. However, the peak in the middle of the graph shows dimers of the adaptors present in the sample.

Increasing the beads:DNA ratio during the clean-up and increasing the number of cycles during the PCR step of the library step did not make any difference. Figures 30-32 represent DNA size and concentration after the modifications of the library prep protocol. They clearly show that changes did not positively impact the DNA yield obtained. DNA yield was insufficient for sequencing. Also, an increase in adaptors and non-desired fragments were obtained.

After engagement with Illumina technical support, libraries were prepared using the Illumina DNA prep kit. Although this kit is not validated for formalin-fixed paraffin-embedded samples, it has been designed for amplicon-based libraries. Figure 33 shows the intensity and the size of the DNA present in the library, showing the expected amplicon size.

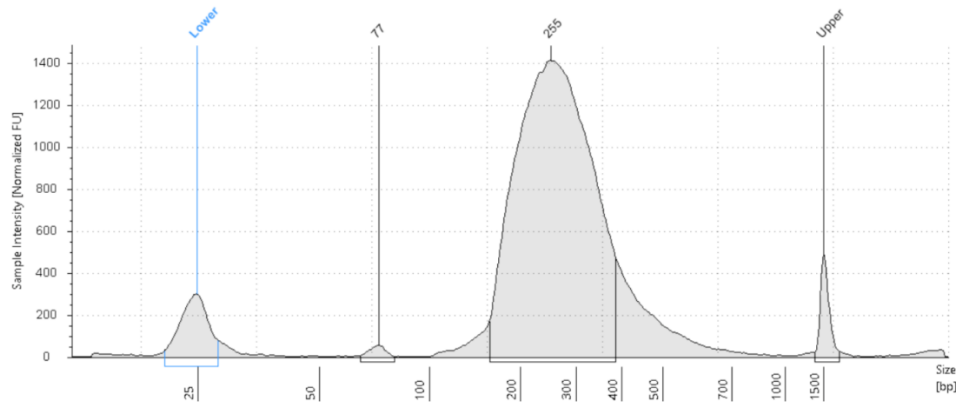


Figure 33. **Tapestation graph representing the sample intensity (FU) for the DNA library's different sizes (bp).** This case shows the pool prepared with the Illumina DNA prep. Peak size perfectly reflects the desired size of the target amplicons.

Due to the issue of DNA yield, all samples tested with the TruSeq kit were repeated using the DNA prep kit for consistency.

### 5.3.5 Validation Edinburgh – Karolinska Anal samples

Twenty-five Fastq files were sent to the International HPV Reference Center in January 2022. A comparison of the HPV 16 sub-lineage for each of the 25 samples was then completed. HPV 16 sub-lineage identified by Karolinska, and this study showed 100% agreement (Table 23).

**Table 23. Comparison of HPV 16 sub-lineage identified by Karolinska Institute.**

<b>ID</b>	<b>HPV 16 sub-lineage identified</b>	<b>Karolinska HPV 16 sub-lineage identified</b>
AC109	A1	A1
AC117	A1	A1
AC118	A1	A1
AC130	A1	A1
AC139	A1	A1
AC140	A1	A1
AC142	A1	A1
AC144	A2	A2
AC146	A1	A1
AC147	A1	A1
AC151	A1	A1
AC152	A1	A1
AC157	A1	A1
AC160	A1	A1
AC161	A1	A1
AC163	A2	A2
AC164	A1	A1
AC165	A2	A2
AC167	A1	A1
AC215	A1	A1
AC21	A1	A1
AC32	A3	A3
AC33	A1	A1
AC9	A1	A1
AC43	A1	A1

#### **5. 4 Discussion**

At the time of writing, this is the first study that has looked into HPV DNA extraction from FFPE samples to provide insight into its ultimate performance in the sequencing of HPV by NGS. Results show that one of the methods, GeneRead, obtained an average higher quantity of DNA, and that no differences between manual DNA Mini Kit and Seegene extraction methods were observed. Notably, however, when DNA



extracts were amplified, no significant difference in the Ct value could be identified between the three assays, which means that initial extraction yield alone is not necessarily a proxy for amplification. Moreover, the results obtained from the designed qPCR have confirmed that the DNA fragment size (potentially caused during FFPE preservation) of nucleic acid extracted from FFPE samples would not be a limitation to performance of short-read sequencing of the HPV 16 genetic material. However, a low concentration of DNA could be a limitation for an NGS direct approach (no enrichment step by PCR or reagents). According to the Illumina DNA kit, the required DNA input ranges from 1 to 500 ng. Data obtained from the three extraction methods shows that the Qiagen DNA mini kit has the most significant number of samples with  $<1 \text{ ng}/\mu\text{l}$  with 14 samples, followed by seven from the Seegene extraction method and two by the GeneRead. Due to the selection of an enrichment NGS method, the importance of DNA concentration is reduced as DNA is subject to several amplifications. Therefore, DNA was obtained from the (automated) Seegene extraction method used for the HPV annotation in anal lesions to perform WGS.

The main limitation of the comparative analysis was that, due to funding constraints, it was impossible to sequence the DNA obtained from the three different extraction methods for each sample, which will be the focus of future work.

Most of the samples received at the SHPVRL for identification/detection of HPV in cancer samples are FFPE specimens, including cervical, anal, and oropharyngeal samples. Up until the current time, FFPE has been used for HPV genotyping with little problems. From time to time, sample results are invalid for testing due to the lack of internal control (beta-globin). The invalid result could be attributed to inhibitors present in the sample that interact during the PCR process and inhibits the DNA amplification. Invalid results have been seen in 0.5% of the cases in the Scottish HPV Reference Laboratory. However, PCR-based testing targets a small region of the L1 gene, potentially overpassing the fragmentation issue. A

short-read sequencing method was selected to avoid any potential fragmentation problem. Moreover, due to the dominance of HPV 16 in anal samples and potential fragment DNA, an approach to enrich only HPV 16 genome was prepared.

Despite the initial problems regarding the low DNA yield with the TruSeq kit, no issues were registered when library prep was performed using the DNA prep kit. The minimum DNA concentration was achieved in every sample. In addition, the comparison performed with Karolinska Institute (100% agreement) verified and reassured that the protocol implemented in the reference laboratory was valid for the HPV 16 sub-lineage identification. The only difference that could explain the unsuccessful use of the TruSeq DNA nano kit is the specimen type. The Swedish study from where the WGS protocol was obtained used cytology samples<sup>52</sup>, while here, it was anal cancer samples preserved in FFPE.

Before starting with any molecular technique is essential to perform a technical optimisation. Even more when the technique has a very high reagent cost. This exercise has enabled me to learn and optimise nucleic acid extraction, library prep and bioinformatic analysis for WGS of HPV. This acquired knowledge will be directly applied to address the research questions about HPV 16 sub-lineages in anal lesions in the next chapter.

## **6. Identification of HPV 16 sub-lineages, association with demographics and clinical variables and overall survival in anal cancer and asymptomatic cohort.**

*Please note that part of the results presented in this chapter have been published: Guerendiain D et al. (2022). Mapping HPV 16 Sub-Lineages in Anal Cancer and Implications for Disease Outcomes.*

### **6.1 Introduction**

HPVs are classified as types based on the nucleotide sequence of the ORF coding for L1. HPV types differ by more than 10% compared with the closest related HPV type<sup>1</sup>. After the carcinogenic evaluation performed by the International Agency for Research on Cancer (IARC), we know that some HPV types have a higher carcinogenic risk than others, and persistent infection with these types could lead to developing pre-cancerous cells and eventually cancer<sup>1</sup>. Moreover, variations within types have led to another classification level, described as lineages (2 - 10% DNA identity) and sub-lineages (0.5% to 2% variation)<sup>2</sup>.

HPV 16 can be divided into four main variant lineages (A/B/C/D) based on genetic variation and 16 sub-lineages based on additional variations within these lineages. As described in Burk et al., 2013 lineage A includes sub-lineages A1-A3 (previously named European) and A4 (Asian); lineage B includes B1 (African-1, Afr1a) and B2 (African-1, Afr1b), B3 and B4. Lineage C includes C1 (African-2, Afr2a), C2, C3 and C4; and D, including D1 (North American, NA1), D2 (Asian-American, AA2), and D3 (Asian-American, AA1) and D4 sub-lineages<sup>15</sup>. In 2001 and later in 2010, Schiffman and colleagues found that non-European sub-lineages were associated with an increased risk of CIN and cervical cancer and that persistence of 2 or more years was linked to the same non-European sub-lineages<sup>201,202</sup>.

Cornet *et al.*, in 2013, identified in cervical cancer that except for sub-Saharan Africa and East Asia, the European variants were prevalent in all regions of the world, whereas the African variant was dominant in the northern sub-Saharan area of Africa and the Asian variant in East Asia<sup>38</sup>. Nicolás-Párraga *et al.*, 2016 explored the HPV 16 lineages and sub-lineages distribution in anogenital cancers from Europe, Asia, and Central/South America. They found that in the cervix, A1-3 was present in 95.6% of the cases in Europe, 78.3% in Central/South America (D in 21.7%) and 80.0% in Asia (12.0% A4 and 7.7% D)<sup>203</sup>.

Sparse information regarding HPV 16 sub-lineages in anal lesions (pre-cancerous and cancer) is available. In the study by Nicolás-Párraga *et al.*, the A1-3 sub-lineages were identified in 96.1% of the European cases, 93.0% in Central/South America (D in 6.9%) and 19.0% in Asia. In this region, A4 was prevalent in 80.9% of the cases<sup>203</sup>. The study conducted by Volpini *et al.* (2017) analysed the HPV 16 variants present in cervical and anal samples. A total of 70.8% (17/24) of the samples were classified as HPV 16 European (E, A lineage) and 29.2% (7/24) as non-European variants<sup>204</sup>. A systematic review by Ferreira *et al.* found that lineage A was present in 100% of anal cancer and 86% of non-tumoral samples. Lineages B and C were present in 2.2% of non-tumoral samples, and lineage D was identified in 9.7% of non-tumoral samples<sup>205</sup>.

Various studies have looked at the association between sub-lineages and cervical cancer risk. Using a WGS assay optimised for HPV genome sequencing, Mirabello *et al.* (2015) examined 3,200 women from a US cohort to assess the correlation between HPV 16 lineages and the risk of precancer/cancer. The study confirmed the previous finding that the B/C/D lineages, taken as a group, had a greater cervical precancer/cancer risk than the A lineage<sup>31</sup>. The researchers carried out a case-control analysis, with the controls consisting of HPV 16-positive women who did not have cervical intraepithelial neoplasia (CIN) grade 2+ after a follow-up period of approximately three years. The analysis demonstrated that the

correlation between HPV 16 lineage and cervical cancer risk differed according to the sub-lineage. In addition, this study confirmed the early observation that some variants present a higher carcinogenic effect in women whose genetic background corresponds to that of the virus.

Between 2017 and 2018, van der Weele *et al.* analysed the whole genome of HPV types 16 and 18 from cervical swabs intending to perform a variant analysis<sup>206,207</sup>. For HPV 16, they analysed the variant diversity and conservation of persistent infections. They could not find an association between sub-lineages and integration and no significant difference in the distribution of A1–3 and A4 variants between CIN and cervical cancer lesions ( $p = 0.936$ ).

Clifford *et al.* (2019) conducted a global sub-lineage analysis of HPV-positive cervical samples. They discovered that sub-lineage A1 was the most extensively distributed worldwide. In contrast, others displayed greater regional specificity (A3 and A4 in East Asia, B1-4 and C1-4 in Africa, D2 in the Americas, and B4, C4, and D4 in North Africa). Additionally, they observed that in regions where A3, A4, and D (sub)lineages were prevalent, there was an elevated risk of cancer compared to the A1 sub-lineage<sup>33</sup>. Furthermore, it has been proposed that some lineages could preferentially associate with different cancer morphologies (i.e., those with squamous or glandular origin). Mirabello *et al.* (2016) found that variants A1/A2 and D2 have a more substantial risk of squamous lesions, while D2/D3 and A4 sub-lineages are strongly associated with glandular lesions<sup>31</sup>. B and C lineages were not associated with adenocarcinomas. Same study looked at the HPV 16 variants at SCC, adenocarcinoma and adenosquamous carcinoma, obtaining similar results<sup>31</sup>. They found that HPV 16 sub-lineages A1-3 were more prevalent in SCC, and HPV 16 D, mainly D3, were increased in glandular cancer lesions.

From 2000 to 2014, most publications determined the variant lineage using phylogenetic parsimony methods based on URR/E6 sequences. However, most publications from 2014 onward, referred to in this introduction, have used next generation sequencing, allowing a complete phylogenetic identification of sub-lineages by looking at the whole genome instead of a smaller region.

Before NGS and whole genome sequencing, different studies investigated the presence of a single nucleotide polymorphism (SNP) associated with persistence and higher risk of progression to cancer<sup>51,208,209</sup>. This SNP has been identified mainly within the European HPV 16 sub-lineages (A1-A3) localised in the E6 gene in position 350, changing from 350T to 350G, resulting in an amino-acid change from leucine to valine. Identification of this SNP could be used as a biomarker to improve diagnosis, patient management, and prognosis.

In addition to sub-lineage identification and cancer risk association, other characteristics/status of the viral genome/status have also been assessed. During the infection of cells by the HPV, integration of part of the HPV genome in the host genome can occur. Data shows that the integration of HPV plays an essential role in overall outcomes. Kim *et al.* 2009, increased E6/E7 oncoproteins as an underlying disruption of E2<sup>20</sup>, resulting in a lower viral replication and poor radiotherapy outcome in patients with low viral load cervical cancers. In head and neck cancers, Vojtechova *et al.*, 2016 found that patients with no detected integration had better survival than integration-positive and HPV<sup>-</sup>patients<sup>21</sup>.

The proportion of cases with HPV integration in anal carcinomas varies between studies, ranging from 32%<sup>210</sup>, 54.8%<sup>211</sup>, to 71%<sup>212</sup>. Moreover, it has been identified differences in integration between HPV types. Lagström *et al.* identified that the proportion of samples with integration was 13% for HPV 16 and 59% for HPV18-positive cervical cancer samples<sup>213</sup>.

### **6.1.2 Aim**

As described in the above paragraphs, several studies have identified the sub-lineages present in the different regions of the world and the risk associated with pre-cancerous and cancerous cervical lesions. However, a limited number of studies with small cohorts have looked at the sub-lineages driven anal lesions (pre-cancerous and cancerous). No study has examined the sub-lineages in anal cancer and analysed the association with demographics, clinical variables, and overall survival. In the UK and Scotland, no information is published on HPV 16 sub-lineage prevalence. Hence, no studies have looked into sub-the association of these sub-lineages with demographic, clinical or survival variables.

Therefore, this chapter aims to identify the HPV 16 sub-lineages in anal cancer samples diagnosed in Scotland and be able to compare those sub-lineages to the ones present in an asymptomatic population. This identification will help identify possible sub-lineage differences between cancers and asymptomatic infections. Moreover, the previous anal cancer chapter has shown that HPV status makes a difference in survival outcomes. Thus, it was required to determine if the sub-lineage status also influences survival outcomes.

From the whole genome sequencing, other information can also be obtained from the samples, including potential missing regions reflecting a potential integration. As described before, integration of the HPV genome in the human genome in oropharyngeal cancers has been associated with a worse prognosis.

Another aim was to identify HPV integration in anal cancers and, if possible, analyse the association with demographics and clinical variables and determine if it is associated with overall survival.

Some single nucleotide polymorphisms (T350) associated with persistent infection have been identified. By using the whole genome data obtained for the sub-lineage identification, it was also aimed to identify the T350 SNP present in both anal cancer and the asymptomatic cohort.

## **6.2 Material and Methods**

### *6.2.1 Sample collection*

#### *6.2.1.1 Anal cancer samples*

HPV 16-positive samples described in Chapter 2 were used for HPV 16 sub-lineage identification. A total of 150 were selected for NGS downstream.

As outlined in Chapter 2, anal biopsy samples were collected between 2009 and 2018 as a routine part of care for treating patients with anal diseases. All biopsy samples were obtained from the South-east of Scotland, representing 3 of 14 territorial health boards in Scotland; NHS Lothian, NHS Borders, and NHS Fife) – these health boards serve a population of 1,396,640 (Data from 2019)<sup>90</sup>, representing 25% of the Scottish population<sup>90</sup>. Samples were archived as formalin fixed paraffin embedded tissue blocks at the Western General Hospital following NHS governance (Human Tissue Governance, (Scotland) Act 2006).

#### *6.2.1.2 Residual rectal swabs from asymptomatic men*

To serve as a control group, residual rectal swabs collected from men attending sexual health clinics without exhibiting anal lesions or symptoms were sequenced for HPV 16 sub-lineage identification. These samples were already tested as part of a study where residual rectal swabs were tested for HPV to assess the prevalence in men who have sex with men (MSM) population (NRS approval SR716). After the study,



samples were archived as nucleic acid at the Royal Infirmary Hospital following NHS governance (Human Tissue Governance, (Scotland) Act 2006).

### *6.2.2 Governance*

Use of anal cancer samples for the current project was approved by the Lothian National Research for Scotland Bioresource (Ref: 20/ES/0061 and application reference SR 1283). Access to residual rectal swab samples was approved by the Lothian National Research for Scotland Bioresource (Ref: 15/ES/0094 and application reference SR1364). The University of St Andrews Teaching and Research Ethics Committee, with reference MD 14482, granted ethical approval to conduct the research.

#### *6.2.3.1 Nucleic acid – anal cancer samples.*

Chapter 5 showed that DNA nucleic acid extracted using the automated Seegene Nimbus (Seoul, Korea) generated material of similar quality and quantity to manual extraction. It was opted to use the nucleic acid obtained in Chapter 4 for HPV genotyping using the Seegene Universal extraction system. DNA was eluted in a volume of 100 µl.

#### *6.2.3.2 Nucleic acid – residual rectal swabs.*

Nucleic acid from residual rectal swabs was obtained using the Qiagen MDx Robot (Qiagen, Germany) and the Seegene Universal extraction kit. The use of 2 different extraction methods was due to the time samples were extracted as the Qiagen instrument was replaced by the Seegene robot. Samples collected between 2018 and 2020 were extracted using the Qiagen platform, while those collected from 2020 were extracted using the Seegene platform. Nucleic acid from both systems was stored at -80°C and -20°C at the Royal Infirmary of Edinburgh.

#### *6.2.4 PCR target enrichment for deep sequencing of HPV 16 – DNA Amplification and primers pooling.*

HPV 16 DNA from samples selected for NGS downstream was amplified using a conventional PCR to amplify the whole genome of HPV 16. Details describing the full protocol are located in the method chapter (Chapter 2).

#### *6.2.5 Library preparation and sequencing.*

The Illumina DNA prep kit was used for library preparation, and sequencing was conducted with the Illumina MiSeq instrument utilising the Illumina MiSeq reagent kit v2 500 cycles. (2x250 bp). The run took 40 hours to complete the 500 cycles.

#### *6.2.6 Bioinformatics pipeline – HPV 16 sub-lineages identification*

An HPV 16 sub-lineage pipeline was not available at the time the analysis was performed, and therefore it was necessary to prepare one. The bioinformatics pipeline has been prepared for UNIX run on a macOS machine. All the different open-source tools used were downloaded from the authors' websites. The raw sequenced reads were mapped to the HPV 16 reference genome [accession number K02718.1] using Burrows-Wheeler alignment (BWA)<sup>94</sup> BCFtools was used for the variant calling and generating a consensus sequence<sup>95</sup>.

The new consensus files were aligned using MAFFT<sup>97</sup> with default parameters, and manual editing was done when necessary. Maximum likelihood trees were created using RAxML<sup>98</sup> v8.2.11 with the GTR substitution model (ML + transfer bootstrap expectation + consensus, one run, 100 reps). The trees generated by RAxML were visualised using Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>). Each

sample was assigned a sub-lineage that corresponded to its nearest neighbour. HPV 16 sub-lineage references were obtained from the Papillomavirus Episteme (PAVE) (<https://www.niaid.nih.gov/research/pave>) for each of the HPV 16 sub-lineages and included in the maximum likelihood tree.

#### *6.2.7 HPV genome integration*

To determine the coverage of the HPV genome of each sample, a visual tool was used (Qualimap<sup>100</sup>). In combination with Artemis<sup>99</sup>, it was possible to identify and annotate regions in the HPV genome that did not contain any reads considered “missing”. Samples with missing reads were repeated from the PCR step to confirm the observations and ensure they were not a result of practical errors. Missing regions were attributed to potential integration in the human genome. Due to the small number of samples with integration, no survival analysis could be performed, but a descriptive assessment of the nature and level of integration was performed.

#### *6.2.8 Association of HPV 16 sub-lineages with demographics and survival outcomes*

To evaluate the relationship between HPV sub-lineages and various factors involving two or more independent variables, a univariate logistic regression analysis approach was used for HPV 16 sub-lineages (HPV 16 A1 positive vs HPV 16-non-A1 positive), including age at diagnosis, collection year, morphology, and health board of diagnosis. The adjustment was made for the age group (<50, 50 – 59, 60 - 69, and 70 or over), sex, response to treatment, stage of cancer, and vital status. The odds ratio (OR) was calculated to measure the strength of the association between HPV 16 sub-lineages and different demographic and clinical data. All statistical analyses were performed using R-studio<sup>214</sup> for macOS.

The study analysed overall survival based on the HPV 16 sub-lineage status (HPV 16 A1 positive vs HPV 16-non-A1 positive) using the Kaplan-Meier method. Cox proportional hazard models were used to derive univariate and multivariate hazard ratios for all-cause mortality based on HPV 16 sub-lineages, adjusting for age (<50, 50-59, 60-69, 70+), sex, stage (I, II, III, IV), and response to treatment (no, yes). All statistical analyses were performed using R-studio (version 1.2.1335). Non-A1 group includes the following sub-lineages: A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3 and D4.

## **6.3 Results**

HPV 16 sub-lineage identification from anal cancers and control cohort through whole genome sequencing

### *6.3.1 Library DNA yield*

#### *6.3.1.1 HPV 16 sub-lineages in the anal cancers*

A total of 119 samples were considered valid for phylogenetic analysis. The sub-lineage of each sample was determined based on its closest genetic neighbour. HPV 16 sub-lineage A1 was identified in 90 anal cancer samples (75.63%), followed by A2, identified in 20 (16.80%) samples. A4 was detected in 5 samples (4.20%). 2 samples were classified as B1 (1.68%), one as A3 (0.84%) and one as D1 (0.84%). HPV 16 sub-lineages in anal cancers are described in depth in Table 24 and Figure 34.

**Table 24. HPV 16 sub-lineages breakdown in the anal cancer cohort.**

<b>HPV 16 Sub-lineage</b>	<b>N</b>	<b>% (N=123)</b>
HPV 16 A1	90	75.63%
HPV 16 A2	20	16.80%
HPV 16 A3	1	0.84%
HPV 16 A4	5	4.20%
HPV 16 B1	2	1.68%
HPV 16 B2	0	0.00%
HPV 16 B3	0	0.00%
HPV 16 B4	0	0.00%
HPV 16 C1	0	0.00%
HPV 16 C2	0	0.00%
HPV 16 C3	0	0.00%
HPV 16 C4	0	0.00%
HPV 16 D1	1	0.84%
HPV 16 D2	0	0.00%
HPV 16 D3	0	0.00%
HPV 16 D4	0	0.00%
<b>Total</b>	<b>119</b>	

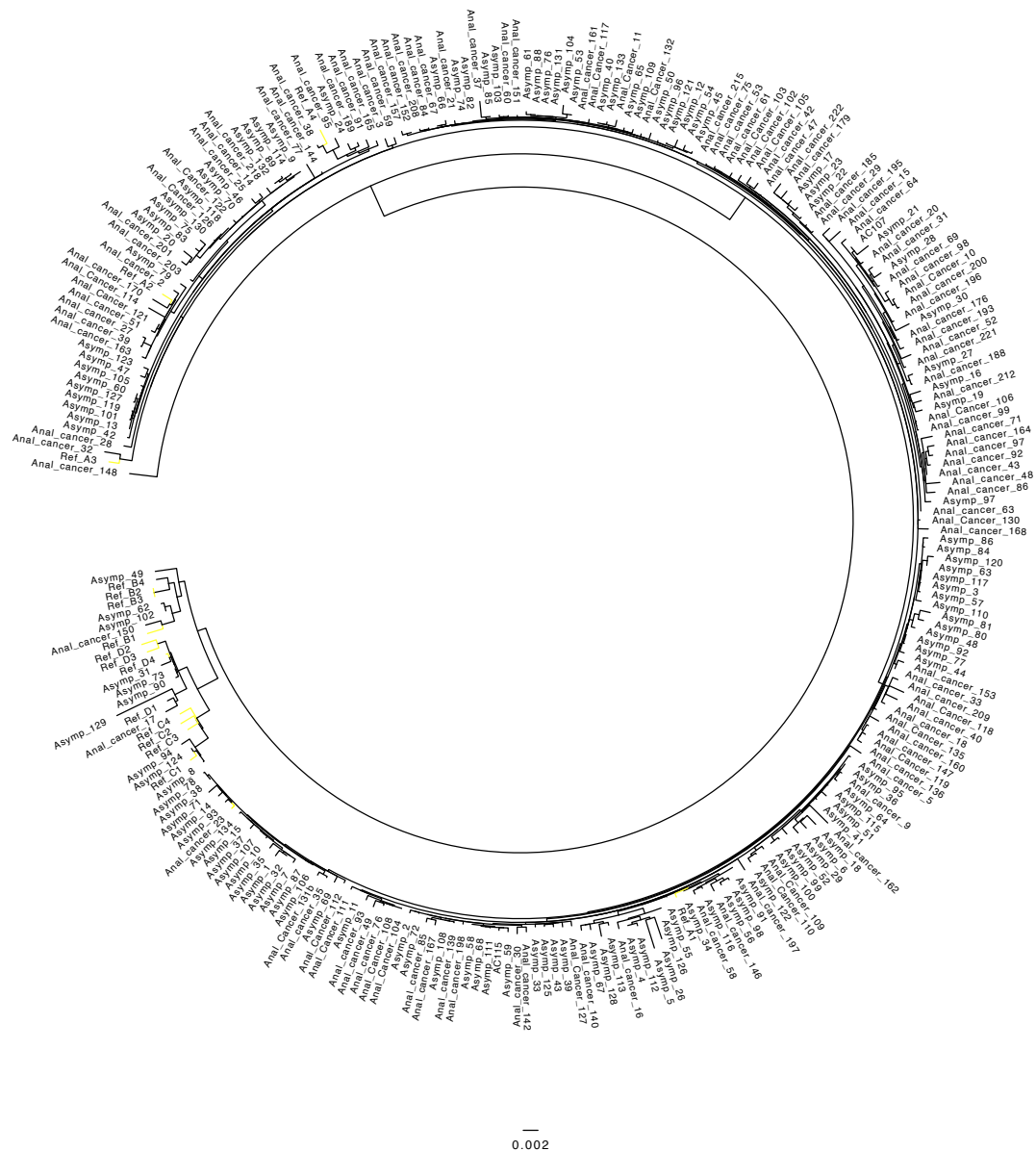


Figure 34. **Phylogenetic tree representing the HPV 16 sub-lineages present in the anal sample cohort based on core SNPs.** Maximum likelihood trees were inferred using RaxML with the GTR substitution model (ML + transfer bootstrap expectation + consensus, one run, 100 reps).

### 6.3.1.2 HPV 16 sub-lineages in the control cohort

From the total 134 control samples considered valid for further phylogeny analysis, most samples were classified (76.12%) as A1, followed by A2, identified in 23 (17.16%) samples. D1 sub-lineage was identified

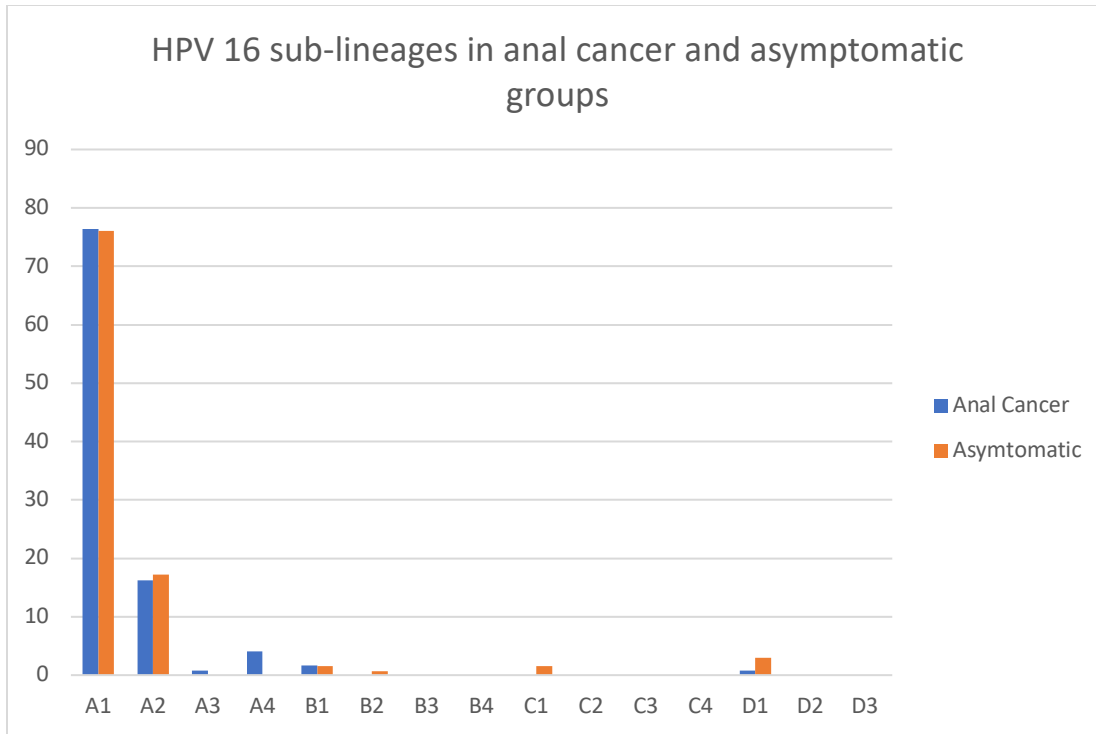
in 4 samples (2.98%), while C1 and B1 were identified in 2 cases each (1.49%). B2 was present in 1 (0.75%). No other HPV 16 sub-lineages could be identified in the control cohort. Table 25 describes the number of cases identified for each sub-lineage, while Figure 35 contains the phylogenetic tree obtained from the control cohort.

**Table 25. HPV 16 sub-lineages breakdown in the asymptomatic population (rectal swabs)**

HPV 16 Sub-lineages Asymptomatic cohort	n = 134	%
HPV 16 A1	102	76%
HPV 16 A2	23	17%
HPV 16 A3	0	0%
HPV 16 A4	0	0%
HPV 16 B1	2	1%
HPV 16 B2	1	1%
HPV 16 B3	0	0%
HPV 16 B4	0	0%
HPV 16 C1	2	1%
HPV 16 C2	0	0%
HPV 16 C3	0	0%
HPV 16 C4	0	0%
HPV 16 D1	4	3%
HPV 16 D2	0	0%
HPV 16 D3	0	0%
HPV 16 D4	0	0%
<b>Total</b>	<b>134</b>	

*6.3.1.3 Differences in prevalence of HPV 16 sub-lineages between anal cancer and control cohort.*

Both cohorts had no differences in the most prevalent sub-lineages (A1 and A2). 75.63% vs 76.0% and 16.80 vs 17%, respectively. However, there were differences in the prevalence of other sub-lineages. A4 sub-lineage was only found in the anal cancer cohort (4.20%), while sub-lineage C1 (1%) and D1 (3%) were only present in the asymptomatic cohort.



**Figure 35. Prevalence (%) of HPV 16 sub-lineages in anal cancer and control cohort.**

*6.3.1.4 Association of HPV 16 sub-lineages with demographic and clinical variables.*

Despite the difference in anal cancer samples with valid sequencing from males (25.21%) and females (74.49%), the prevalence of A1 and A2 is almost identical in both groups (Table 26). The only differences are the A3 presence in males (1/30), higher prevalence of A4 (4/89) and the absence of B1 and D1 in males but present in the female cohort.



**Table 26. HPV 16 sub-lineage distribution in males and females in the anal cancer cohort.**

Anal cancer HPV 16 sublineages	Males		Females		Total
	n	%	n	%	
<b>A1</b>	23	76.67%	67	75.28%	90
<b>A2</b>	5	16.67%	15	16.85%	20
<b>A3</b>	1	3.33%	0	0.00%	1
<b>A4</b>	1	3.33%	4	4.49%	5
<b>B1</b>	0	0.00%	2	2.25%	2
<b>D1</b>	0	0.00%	1	1.12%	1
<b>Total</b>	30		89		119

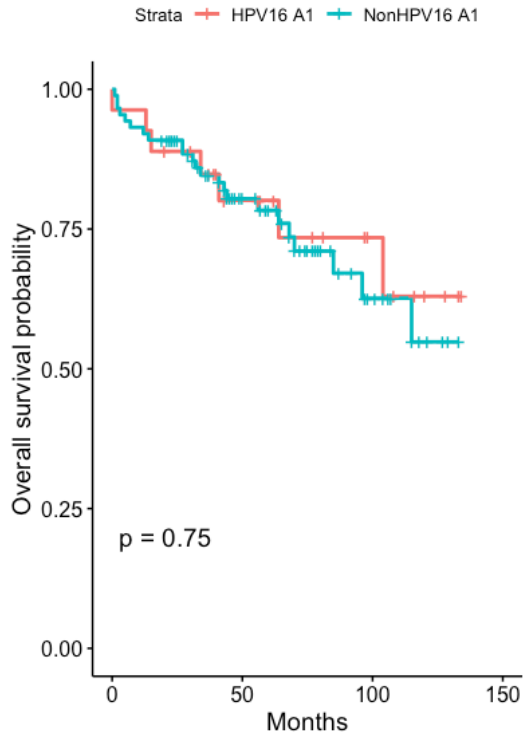
Four of the 119 samples did not contain vital status information and therefore have not been included in the analysis. (4 = 2 A1, 1 B1, 1 A2). Therefore, 115 samples were included in the HPV 16 sub-lineage association with demographic and clinical variables analysis. Due to the dominance of A1 in the anal cancer cohort and the low number of other sub-lineages present, logistic analysis and odds ratio analysis were performed based on HPV 16 sub-lineage A1 positive or A1 negative. No significant differences in the univariate association of A1 positivity with sex, age, response to treatment, stage, and vital status were observed. Adjusted analysis by sex, age, response to treatment, stage and vital status shows no significant differences for any of the demographics or clinical variables (Table 27)).

**Table 27. A1 positivity status according to demographics and clinical variables.** The odds ratio (univariate and adjusted) was calculated for A1 positivity status.

Variable	Level	Unadjusted OR (95% Cis)	p-value	Adjusted OR (95% Cis)	p-value
<b>Sex</b>	<b>Male</b>	1		1	
	<b>Female</b>	1.12 (0.39 - 2.92)	0.827	1.09 (0.37 - 3.00)	0.87
<b>Age</b>	<b>&lt;50</b>	1		1	
	<b>50 - 59</b>	1.20 (0.27 - 4.80)	0.8	1.11 (0.24 - 4.56)	0.89
	<b>60 - 69</b>	1.50 (0.34 - 5.93)	0.57	1.82 (0.40 - 7.67)	0.416
	<b>70 and over</b>	1.37 (0.30 - 5.67)	0.666	1.63 (0.34 - 7.41)	0.529
<b>Response to treatment</b>	<b>No</b>	1		1	
	<b>Yes</b>	1.02 (0.26 - 3.26)	0.968	1.18 (0.34 - 7.41)	0.528
<b>Stage</b>	<b>I</b>	1		1	
	<b>II</b>	1.36 (0.37 - 4.62)	0.625	1.28 (0.33 - 4.51)	0.706
	<b>III</b>	1.60 (0.40 - 6.11)	0.486	1.56 (0.38 - 6.06)	0.522
	<b>IV</b>	1.80 (0.36 - 10.40)	0.478	3.03 (0.42 - 29.47)	0.289
<b>Vital Status</b>	<b>Alive</b>	1		1	
	<b>Deceased</b>	1.01 (0.39 - 2.85)	0.983	0.92 (0.25 - 3.81)	0.907

#### 6.3.1.5 HPV 16 sub-lineages and overall survival

The Kaplan-Meier estimator calculated overall survival by categorising HPV 16 sub-lineages into two binary groups: A1 positive and non-A1 positive sub-lineages. No differences in overall survival were found between both sub-lineage groups ( $p=0.75$ ) (Figure 36).



**Figure 36. Kaplan-Meier survival curve stratified by HPV 16 sub-lineages (A1 vs non-A1).** Survival time is expressed in months from the diagnosis date. Data was censored on 31st July 2020.

Table 28 presents the overall survival results based on HPV 16 sub-lineages stratified by various clinical and demographic factors such as age group, sex, cancer stage, and response to treatment. The sub-lineages were categorised into two groups: A1 and non-A1, with A1 being the reference group. The univariate analysis showed no significant association between non-A1 sub-lineages and improved overall survival, with a hazard ratio (HR) of 0.87 (0.37 – 2, p=0.751) compared to A1. On the other hand, stage IV vs stage I and no response to treatment vs response were significantly associated with worse overall survival in the univariate model with HRs of 15.7 (3.38 – 72.8), p<0.001 and 0.11 (0.05 - 0.25) p<0.001, respectively. However, after adjusting for age, gender, stage, and response to treatment, non-A1 sub-lineages did not significantly influence overall survival compared to A1, with an HR of 0.83 (0.28 – 2.46, p=0.743).

**Table 28. A univariate and multivariate hazard ratio of HPV 16 sub-lineages derived using Cox regression.**

Variable	Level	Unadjusted HR (95% Cis)	p-value	Adjusted HR (95% Cis)	p-value
HPV 16 Sub-lineage	A1	1		1	
	Non-A1	0.87 (0.37 - 2)	0.751	0.83 (0.28 - 2.46)	0.743
Sex	Male	1		1	
	Female	1.2 (0.48 - 2.9)	0.71	0.88 (0.32 - 2.39)	0.795
Age	<50	1		1	
	50 - 59	1.10 (0.33 - 3.70)	0.877	0.83 (0.21 - 3.26)	0.788
	60 - 69	0.85 (0.26 - 2.8)	0.795	2.67 (0.607 - 11.72)	0.194
	70 and over	1.54 (0.48 - 5.0)	0.466	5.56 (1.082 - 28.58)	0.04
Stage	I	1		1	
	II	1.7 (0.37 - 8.1)	0.49	2.34 (0.47 - 11.74)	0.302
	III	2.4 (0.50 - 11.6)	0.274	2.26 (0.42 - 12.27)	0.344
	IV	15.7 (3.38 - 72.8)	<0.001	15.95 (2.45 - 103.82)	0.004
Response to treatment	No	1		1	
	Yes	0.11 (0.05 - 0.25)	<0.001	0.12 (0.03 - 0.39)	<0.001

When considering only A1 and A2 samples, there were no significant differences in the hazard ratio (HR) found when using A1 as the reference (HR 0.74, 0.25 – 2.1, p=0.575).

#### 6.1.3.6 Integration anal cancer

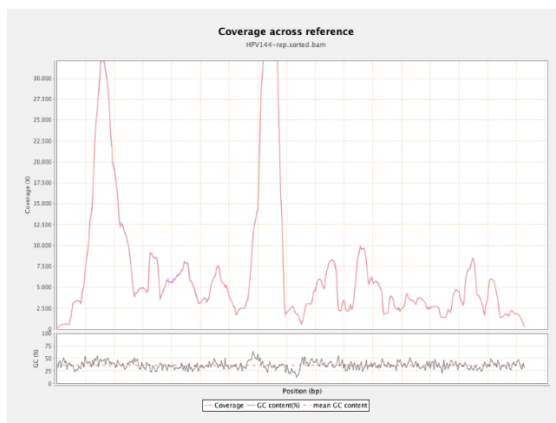
Using NGS, we should obtain hundreds or thousands of reads in every position of the target DNA. However, sometimes no reads are obtained for some regions. Although not the study's primary objective, 10.92% (13/119) of the anal cancer samples had missing parts of the HPV 16 genome, indicating the

potential integration of HPV 16 into the human genome. The missing region was confirmed through repeat sequencing. The E2 gene was the most commonly absent region, followed by E4, E5, L2, E1, and L1 (Table 29). However, all cases retained the E6 and E7 oncogenes. There was no evidence of integration found in the asymptomatic cohort.

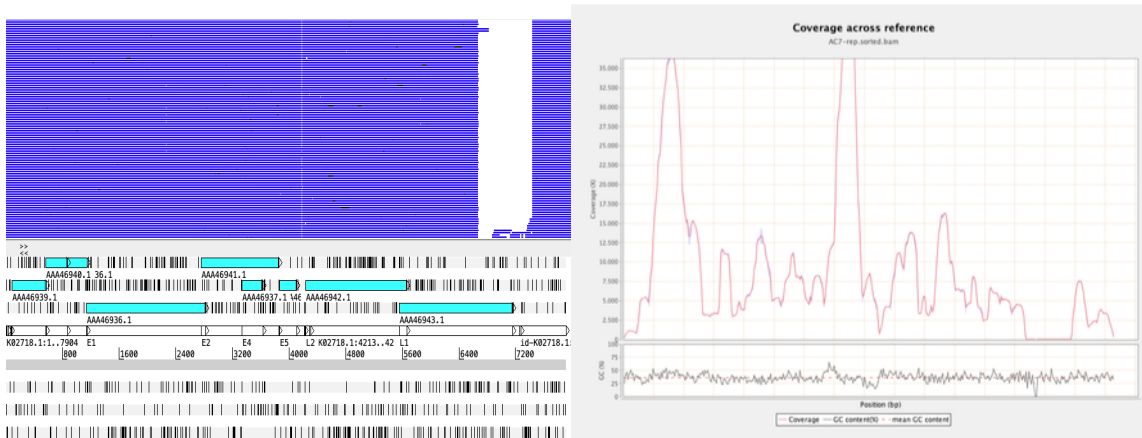
**Table 29. Missing genes in the suspect samples of HPV being integrated.**

HPV 16 Genes	Missing genes (even partially)
E6	0
E7	0
E1	7
E2	11
E4	10
E5	10
L2	10
L1	7

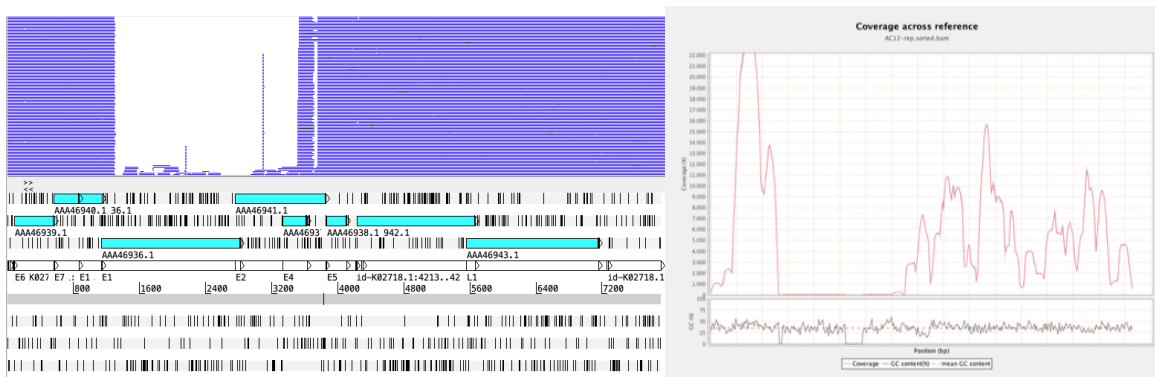
Figures 37 to 50: Reads coverage obtained from different anal cancer samples. Figure 37 shows a full coverage of the HPV genome, while the rest show no reads in specific regions. This absence of reads in different locations of the HPV genome is associated with the potential integration of the HPV genome in the host during replication.



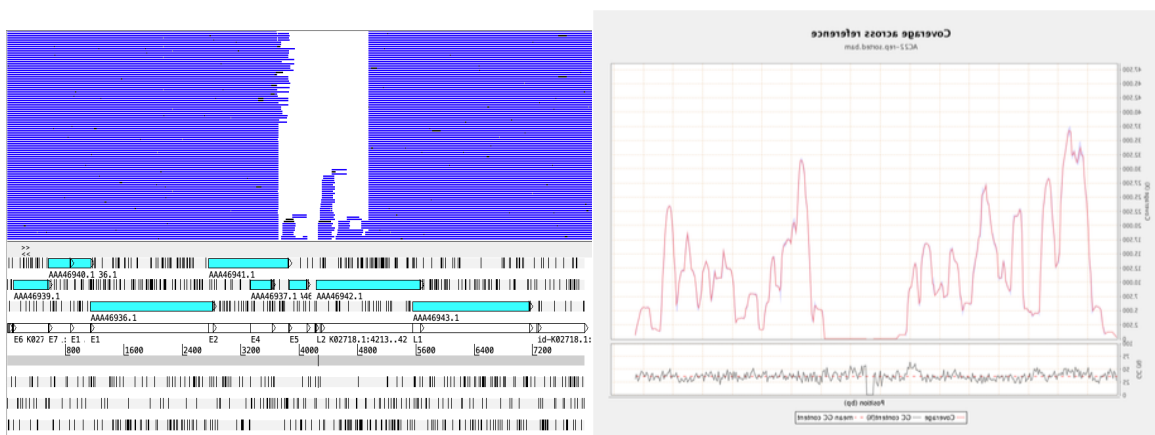
**Figure 37. Qualimap Plot showing an anal cancer sample with no integration. Mapped paired reads 298,541 / 99,72%.**



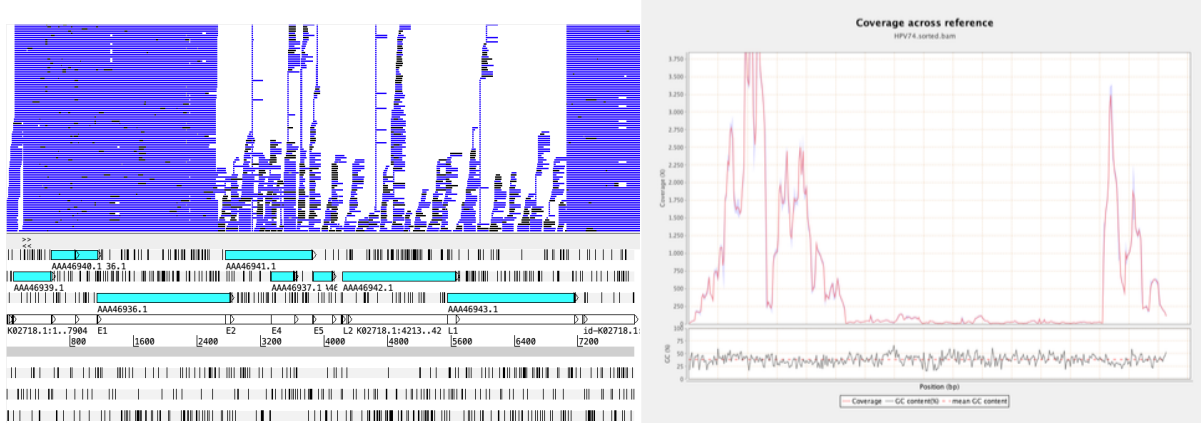
**Figure 38. Potential Integration of E1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



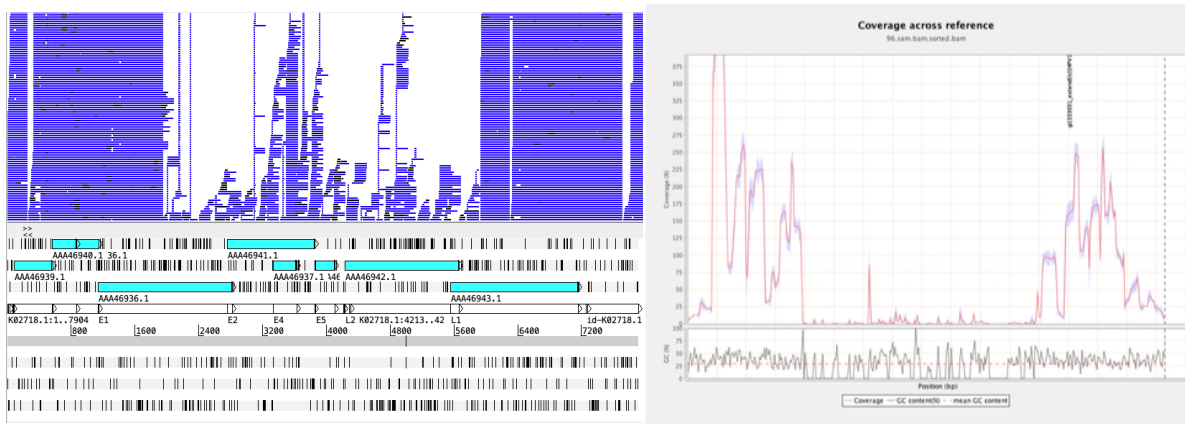
**Figure 39. Potential integration of HPV E1, E2 and E4.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



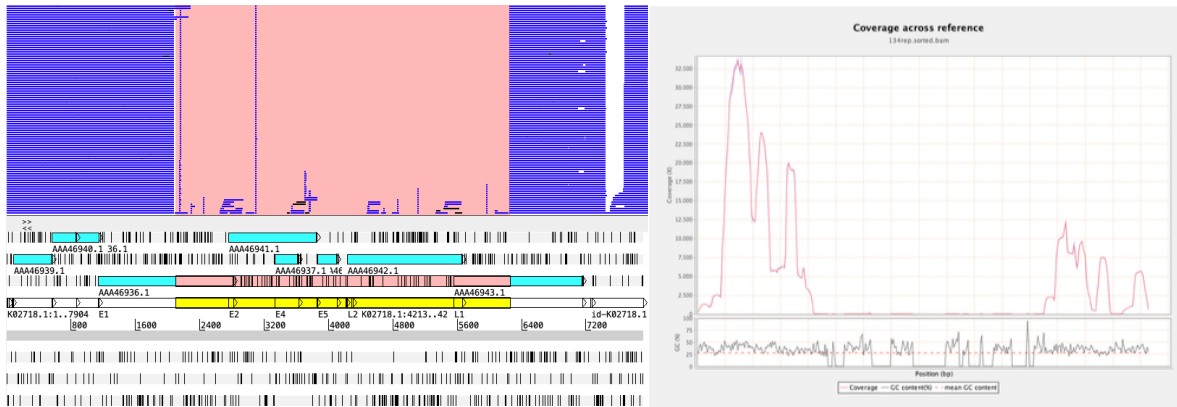
**Figure 40. Sample AC22: Plausible integration of part of E2, E5 and part of L2.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



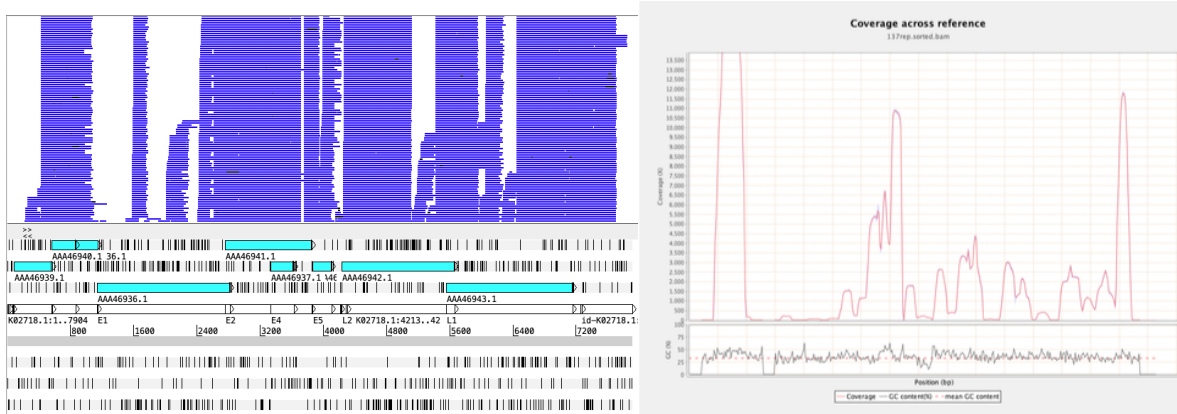
**Figure 41. Sample AC74: Plausible integration of genes E2, E4, E5, L2 and L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



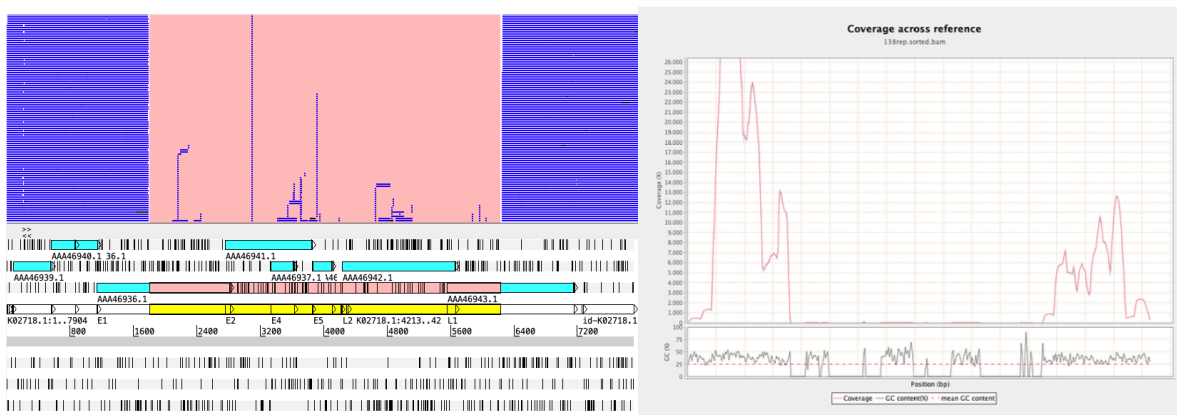
**Figure 42. AC96: Plausible integration of: Part of E1, E2, E4, E5, L2 and first part of L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



**Figure 43. AC134: Plausible integration of part of E1, E2, E4, E5, L2 and first part of L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).

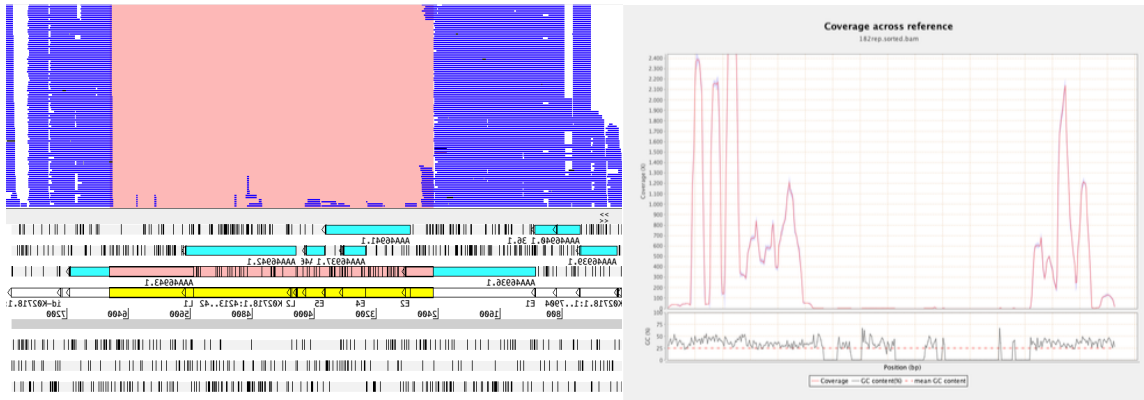


**Figure 44. AC137: Plausible integration of part of E1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).

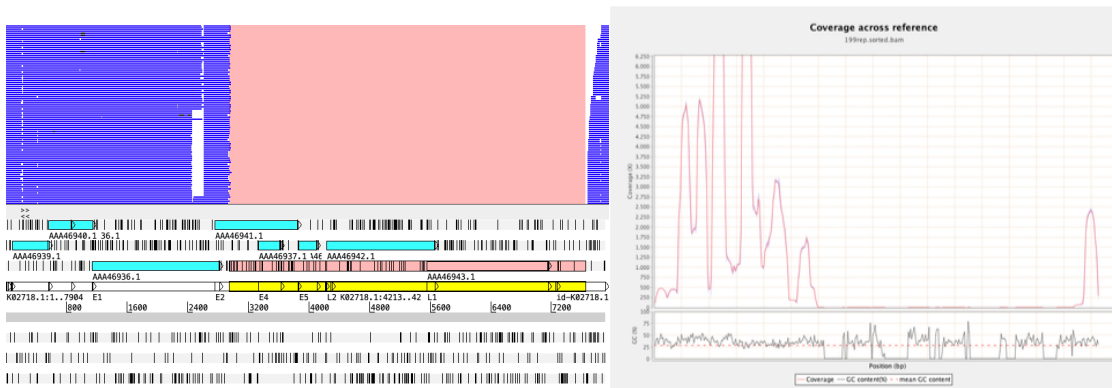


**Figure 45. AC138: Plausible integration of part of E1 E2, E4, E5, L2 and part of L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).

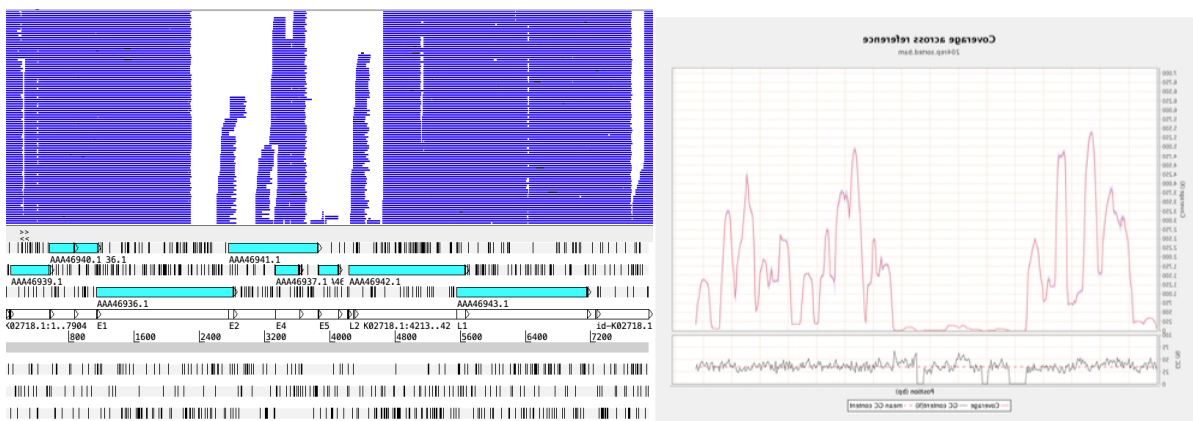




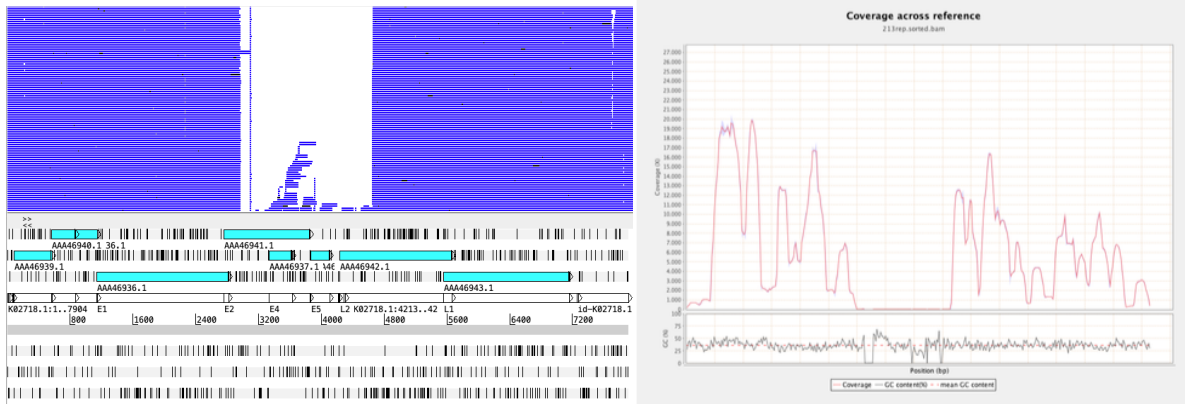
**Figure 46. AC182: Plausible integration of part of E1, E2, E4, E5, L2 and part of L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



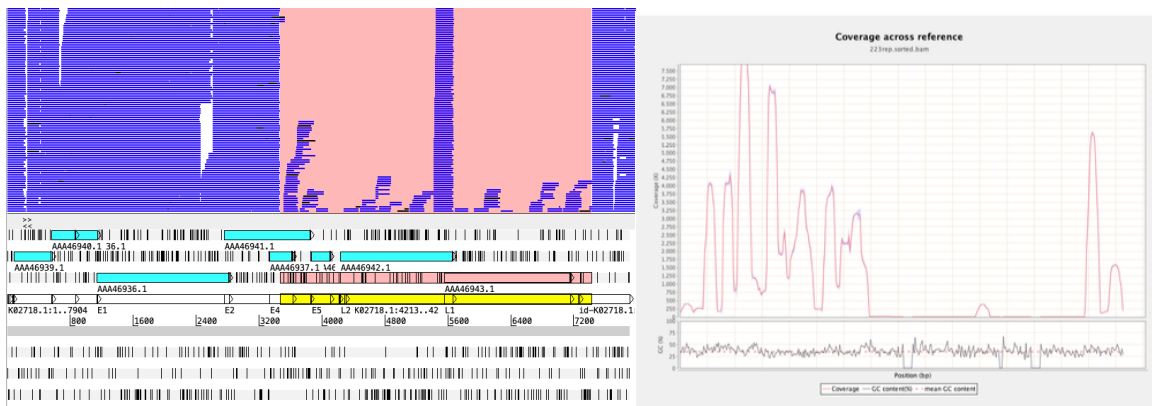
**Figure 47. AC199: Plausible integration of part of E2, E4, E5, L2 and L1.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



**Figure 48. AC204: Plausible integration of part of E1, E2, E4, E5, and part of L2.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



**Figure 49. AC213: Plausible integration of part of E2, E4, E5, and part of L2.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).



**Figure 50. AC223: Plausible integration of part of E2, E4, E5, L1 and L2.** Artemis displaying the bam file and read coverage (left) and Qualimap plot of total read coverage (right).

As described in Table 30, out of the 13 samples with a potential integration, 8/13 were females between 36 and 61 years old (average 48.4). 5/13 samples were from males between 56 and 73 years old (average 64 y/o). Four patients were at stage IV, four at stage II and one at stages III and I. Unfortunately, two samples did not contain stage information. 8/13 patients were deceased when data was censored, four were alive, and 1 had no vital status. Using the viral load information obtained for the analysis described in Chapter 2, 5/13 of the samples had a low viral load, 1 had a medium viral load, and 7 had no valid viral load information.

The target used for the ddPCR viral load analysis was L1. Those samples that did not have a valid viral load have a potential integration of the whole or partial L1 gene.

**Table 30. Demographics of anal cancers with a plausible integration.**

Sample ID	Age	Sex	Stage	VL (ddPCR)	VL (qualitative)	Vital status
AC 7	41	Female	IV	30.3	Medium	Deceased
AC12	60	Female	IV	3.6	Low	Deceased
AC22	56	Male	IV	1.93	Low	Deceased
AC74	65	Male	IV	NA	NA	Deceased
AC96	62	Male	II	NA	NA	Deceased
AC134	61	Female	II	NA	NA	Alive
AC137	64	Male	II	4.1	Low	Alive
AC128	73	Male	II	NA	NA	Deceased
AC182	56	Female	III	NA	NA	Deceased
AC199	45	Female	IV	NA	NA	Alive
AC204	37	Female	NA	NA	NA	NA
AC213	52	Female	NA	3.7	Low	Deceased
AC223	36	Female	I	1	Low	Alive

#### 6.3.1.7 SNPs - 350T – 350G

The nucleotide in the E6 gene of every anal cancer sample was analysed. A total of 52/115 (45.2%) anal cancer samples had thymine (350T), while 63/115 (54.8%) had guanine (350G). Univariate Cox analysis showed no difference in survival of 350G when 350T was used as a reference (0.91, 0.44- 1.9, p=0.794). When adjusted by age group, sex, stage and response to treatment, HR for 350G was 1.06 (0.404 – 2.77, p=0.908) when using 350T as a reference.

## 6.4 Discussion

As described in chapter 4 of this thesis (Anal disease in the South-East of Scotland. HPV prevalence, association with demographics and survival), 93.3% of anal cancer diagnosed in Scotland between 2019 and 2018 were caused by HPV 16. Among these, 76% of cases belonged to the A1 sub-lineage, followed

by A2 (16%), based on whole genome sequencing. The prevalence of A1 and A2 in the asymptomatic male control group was found to be similar. However, a few discrepancies were noted; sub-lineage A4 was present in the anal cancers (4.2%), while this was absent in the control group; the presence of the C1 sub-lineage was only detected in the control group; and sub-lineage D1 presence was higher in the control group (3%) than in the anal cancer cohort (0.84). The high prevalence of A1 and A2 sub-lineages is in line with previous European studies such as Gonçalves *et al.* (2022), who found lineage A (mainly A1) to be the more prevalent in the anal canal of asymptomatic men<sup>215</sup>, and Nicolás-Párraga *et al.* (2016) who found that A1–3 sub- in 96.1% of European cases<sup>203</sup>. Beyond Europe, Volpini *et al.* (2017) conducted a study in Brazil that examined HPV 16 variants in cervical and anal samples. They found that a lower proportion of anal cancer samples (70.8%) were classified as A1-3 sub-lineages compared to the results observed in European studies<sup>204</sup>. In addition, Gonçalves *et al.* (2022) also found a higher prevalence of non-A lineages in MSM with transient infection<sup>215</sup>. This study has detected a higher proportion of non-A lineages in the asymptomatic MSM cohort. However, a small number of cases require further studies to confirm this.

The findings from the study contribute to the scarce knowledge of the distribution and consequences of HPV sub-lineages in the anus. Although our results did not indicate strong correlations with demographics and underlying disease status, further research with more significant participant numbers is necessary to validate or reject these conclusions.

As far as I know, no other studies have examined the relationship between HPV 16 sub-lineages and survival in anal cancer patients. The analysis presented in this chapter did not show an impact on overall survival when comparing A1 and non-A1 sub-lineages in the univariate and adjusted analysis. A recent study by Lang Kuhs *et al.* (2022) analysed the connection between the genetic variation of HPV 16 and

clinical outcomes in patients with HPV 16-positive oropharyngeal cancer<sup>216</sup>. They discovered that patients with one or more high-risk single nucleotide polymorphisms (SNPs) had shorter median survival times. Many of these SNPs were associated with the D2 sub-lineage, also linked to an increased risk of cervical cancer<sup>31</sup>. Unfortunately, I could not identify any cases of the D2 sub-lineage, so I was unable to investigate this connection further. Identifying other high-risk SNPs in other sub-lineages could greatly aid in patient and treatment management.

This study has certain limitations that must be noted. The asymptomatic population only consisted of men, while the cancer population had more women (75.63%) than men (24.37%). This disparity was due to practical considerations with the available sample material. Despite this, the data showed no differences in the distribution of HPV 16 sub-lineages between men and women in the anal cancer group. As previously mentioned, confirming these findings with a larger sample size would be valuable. Although the number of cancer cases in this study was not insignificant (n=253), considering the Scottish European age-standardised rate (EASR) for anal cancer is 2.6 per 100,000 person-years at risk in 2017, larger sample sizes may be necessary for detecting rarer sub-lineages with more accuracy.

Currently, there is no anal cancer screening program in the UK. Nevertheless, an opportunistic vaccination program for men who have sex with men has been in place since 2017, and the national HPV vaccination program has been gender-neutral since 2019. A study by Godi *et al.* found that HPV 16 lineages B, C, and D showed slightly reduced sensitivity (<2-fold) to the nonavalent vaccine compared to lineage A<sup>217</sup>. The high prevalence of lineage A in this study's samples could be positive for vaccine efficacy, especially with the implementation of gender-neutral vaccination in the UK and several other countries.

This study has demonstrated that it is technically feasible to detect HPV 16 sub-lineages in both anal cancer samples and residual material from rectal swabs in an asymptomatic population. While minor

differences in the prevalence of non-A sub-lineages were observed between cancer and asymptomatic populations, the low presence of these sub-lineages made it impossible to complete a full investigation. It would benefit from further study to understand their significance and implications. The dominance of lineage A is in line with previous European data and suggests that sub-lineage identification may not be a reliable predictor of prognosis in anal cancer. However, ethnicity and diversity of the population analysed seem to play an important role in these studies; as shown by Brim *et al.*, 2019, non-A1 sub-lineages were significantly associated with cancer among African Americans <sup>218</sup>. Therefore, in future studies would also be good to get ethnicity information to perform a complete analysis and obtain a better understanding of the influence of these sub-lineages on cancer risk and overall survival.

## 7. Applications and translation of next generation sequencing

### 7.1 Introduction

The detection of human papillomavirus (HPV) infection has advanced significantly over the last two to three decades. Initial low-throughput hybridisation/blotting techniques preceded broad-spectrum signal amplification assays, which were then replaced by rapid high-throughput target-amplification assays involving quantitative polymerase chain reaction (qPCR)<sup>219</sup>. Amplification tests can detect individual genotypes (or a group thereof) and have become the central pillar of HPV-based screening and clinical diagnosis<sup>220</sup>. In the last ten years next generation sequencing (NGS) has erupted in the microbiology molecular diagnostic field due to the reduction of cost, increased knowledge of genomics and a greater understanding of what can be achieved from the data obtained. NGS allows deep sequencing of samples, generating millions of reads in parallel to determine entire genomes, or can be used to focus on specific genome regions.

For HPV detection purposes, NGS can go beyond the simple detection of HPV and provide insight into the likely course and clinical consequences of the infection<sup>220</sup>. The technology used for HPV detection in results chapters 3 and 4 were PCR based, and HPV was discerned at the HPV “type” level using assays that focussed on one gene target only. NGS and a whole genome analysis approach were used for sub-lineage identification in anal samples (as described in Chapter 6). The NGS approach permitted in-depth coverage of all the HPV 16 genome and enabled detailed analysis of its variability and association with underlying disease status and clinical outcomes. Additionally, thanks to the data obtained through WGS, it was possible to identify the potential integration of the HPV virus in the human genome and the prevalence of SNPs at position 350 for the E6 gene associated with higher persistence.

NGS improves the sensitivity achieved by standard PCR-based tests and can detect novel types and known types that are distantly related to primers/probes, which may escape detection using standard molecular approaches<sup>221–223</sup>. However, given the current availability of PCR-based assays that show validated performance for specific applications (including screening), the relatively high cost of NGS and the unknown magnitude of clinical benefits conferred by NGS for HPV detection in the past, routine implementation of this technology in clinical/service laboratories has been limited. It is more likely to be present in specialist/reference centres.

Thanks to whole genome sequencing and NGS technology experience and skills obtained during the preparation of this thesis, there was an opportunity to produce a chapter that covered how to set up a next generation sequencing approach for HPV diagnosis, what different options exist and how it may be possible to integrate it into a clinical laboratory service such as within a reference laboratory where much of the practical work for this thesis was performed. This chapter will consider potential NGS applications for HPV diagnosis and compare them with the routine assays used in an existing HPV reference laboratory. A case study will be presented describing the best approaches to perform HPV sequencing using already available sequencing capacity.

#### *7.1.1 Current HPV tests used for clinical diagnosis.*

There are multiple molecular tests available for the detection of HPV<sup>50</sup>. While these molecular tests vary in the approach/technology used, most commercial tests are based on HPV amplification and detection of the HPV nucleic acid (DNA or mRNA). Other tests use a hybridisation approach, where HPV is captured, resulting in fluorescence emission or colour change.



Due to the high sensitivity, specificity, and robustness, most of the commercial HPV tests available are PCR based. Some of these assays have been developed in commercial laboratories and then validated through the execution of a number of analytical and clinical studies. For HPV tests with a (cervical) screening application, clinical validation is achieved through comparative performance with a gold standard assay (Qiagen Hybrid capture)<sup>224</sup>. In contrast, in-house tests are those distilled from other published methods or developed by the laboratory. Due to the increasing and detailed requirements of external accreditation organisations (e.g., UKAS in the UK) that work to international quality standards (such as ISO15189), in-house tests have been reduced in laboratories with a service remit. However, they still exist and can form an essential part of the repertoire, particularly in specialist/reference environments.

As in other laboratory medicine fields, test requirements are sometimes different; they can vary depending on the number of samples required to be processed, turnaround times etc. What is required for a large screening laboratory may not suit other labs, such as a reference lab, where more specialist tests are often offered.

#### *7.1.2 Next generation sequencing*

Although NGS has been used in HPV research for some years, NGS in screening and diagnosis service laboratories is not yet extensively used. High cost per sample, specific laboratory infrastructure and training requirements, and the associated significant capital investment which is needed may have slowed down the adoption of this technology. However, investment by national governments in NGS due to the COVID-19 pandemic, and the experience acquired during the same period in the use of genomic data for clinical and public health benefit, have positively impacted the appreciation of NGS's potential and how it ultimately may improve the patient pathway and epidemiological precision. This progress could be

translated into the routine detection of other pathogens, like HPV. The advancements made by SARS-CoV-2 sequencing provide an opportunity to reflect on how this may improve HPV testing.

A PubMed search (March 2023) was performed with the following terms ((next generation sequencing) OR (NGS)).

Table 31 describes the number of publications of NGS + HPV and the number of publications where NGS was used. It can be observed that the number of publications on HPV and NGS has increased over the last ten years, from 3 in 2011 to 95 in 2021. However, when we compare the number of these publications with the total number of studies published in PubMed in 2021 (where NGS or next generation sequencing was performed), only 0.6% of those were HPV related.

**Table 31. The number of results obtained in PubMed.** Search performed in PubMed on the 2<sup>nd</sup> of March 2023. \*Complete search = (((Human Papillomavirus) OR (HPV)) AND ((next generation sequencing) OR (NGS))). †Complete search = (Next generation sequencing) OR (NGS)

Search query:	HPV + NGS*	Next generation sequencing OR NGS †	%
<b>2022</b>	59	12228	0.4%
<b>2021</b>	95	14680	0.6%
<b>2020</b>	85	13982	0.6%
<b>2019</b>	75	12025	0.6%
<b>2018</b>	62	10287	0.6%
<b>2017</b>	48	9298	0.5%
<b>2016</b>	37	8922	0.4%
<b>2015</b>	44	7643	0.6%
<b>2014</b>	23	5936	0.4%
<b>2013</b>	20	4373	0.5%
<b>2012</b>	14	3130	0.4%
<b>2011</b>	3	1907	0.2%

### 7.1.3 Application of NGS for HPV

NGS improves conventional diagnostic tests and provides opportunities to understand the infection better, with a potential for more specific prognostication. For HPV in particular NGS may add value to the following applications:

- **Genotyping:** target sequencing focus only on L1, E6/E7 or WGS. Identification of no conventional HPV types or novel types.
- **Variant & sub-lineage identification:** by performing phylo-genomic analysis.
- **Methylation:** NGS can help identifying HPV DNA methylation. Data indicated that HPV 16 CpG methylation at L1 and L2 sites (for example) is a biomarker of cervical precancer.
- **Integration:** HPV integration in the human genome by looking at missing regions of the HPV genome.
- **Viral load:** As identified previously, viral load may play a role in the overall survival of HPV associated cancers.
- **SNP identification:** NGS can help determine SNPs with precision that may be prognostic.
- **HPV circulating DNA (ctDNA):** Detection of HPV or tumour DNA in blood can help assess the treatment response.

### 7.1.2 Aim

This chapter aims to consider and present the potential applications of NGS in HPV testing repertoire within a service/reference laboratory context. In so doing, I will consider the different NGS options. Furthermore, it will also produce a roadmap for incorporating an NGS approach into a service repertoire.

## 7.2 Material and Methods

In this chapter, different NGS applications for HPV diagnosis will be presented and compare the potential advantages and disadvantages concerning existing HPV tests for:

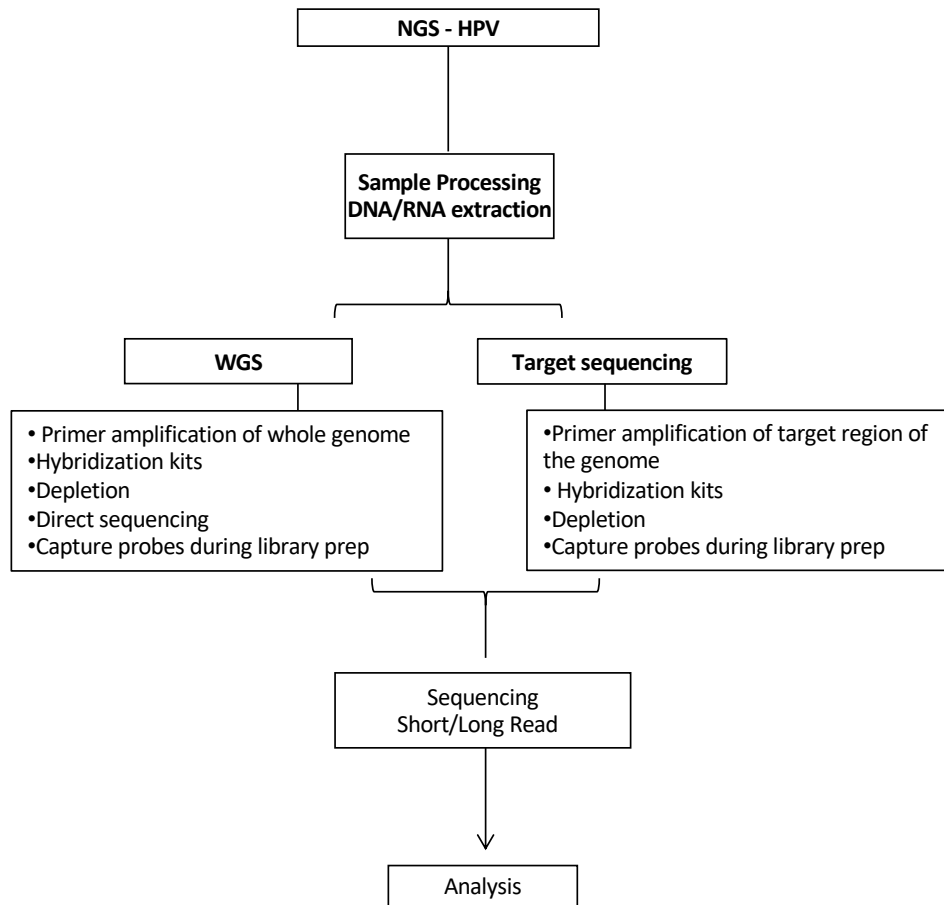
- Cervical screening/disease management.
- Immunisation surveillance.
- Annotation of cancers for HPV status.
- Deep interrogation of the HPV genome.
- Research and development, including biomarker design and applications.
- Liquid biopsy detection (Comparison with ddPCR).

## **7.3 Results**

### *7.3.1 NGS assays for HPV sequencing - an overview.*

As described previously, NGS allows not only the identification of different types of HPV present in a sample but allows deep interrogation of the HPV genome, permitting the identification of new types, integration for parts of the viral genome in the host or detection of new lineages.

As described in Figure 51, two different approaches that use massively parallel sequencing technology exist. One is the WGS approach, where the entire genome is sequenced, and the other is by targeting a region of the genome. Both would require similar protocols; however, WGS adds extra complexity to the downstream analysis as a larger volume of fragmented sequence need to be processed.



**Figure 51. Diagram of the HPV NGS workflow for whole genome sequencing or target sequencing.**

### 7.3.1.1 NGS as a genotyping tool.

We can sequence the virus's entire genome using NGS or focus only on smaller regions/parts. Conventional HPV detection/genotyping has been based on specific viral genome regions like L1, E6 or E7. If we sequence and analyse these regions, we can also identify the HPV type present in the sample and use NGS as a genotyping tool. As we are only sequencing small parts, the number of samples we test per sequencing run would be higher than when performing whole genome sequencing. Additionally, the number of samples per run would be limited by the number of indexes we can use. For example, we could sequence 96 samples in a MiSeq run, but it could go up to 384 if using a NextSeq. Some publications have

looked into using NGS as a genotyping tool, like Yi *et al.*, 2014, where using an IonTorrent platform, they managed to genotype over 1000 samples in one run<sup>54</sup>.

Table 32 describes the cost of a genotyping approach using a MiSeq platform and Illumina reagents for 96 samples. Sequencing only specific targets (L1) for genotyping, the cost associated with sequencing 96 samples was £36.93 (Table 32). The cost includes extraction, PCR amplification, library prep (including Qubit costs) and sequencing.

Due to the length of the sequencing process, hands on time, and high cost, using NGS as a genotyping tool is not cost-effective; when we have other tests that can provide the same information quicker and cheaper. Conventional HPV tests can obtain a result in less than 8 hours (including nucleic acid extraction).

**Table 32. The cost associated with an NGS genotyping approach (not WGS).**

<b>Main reagents</b>					
<b>Supplier</b>	<b>Item</b>	<b>Reference</b>	<b>Unit Price</b>	<b>n samples</b>	
Qiagen	Qiagen Multiplex PCR Kit (100)	206143	£179.00	48	£3.73
Illumina	Illumina® DNA Prep, (M) Tagmentation (96 Samples)	20018704	£1,836.29	96	£19.13
Illumina	Nextera™ DNA CD Indexes (96 Indexes, 96 Samples)	20018705	£169.82	96	£1.77
Illumina	MiSeq Reagent Kit v2 (500cycle)	MS-102-2003	£748.90	96	£7.80
Qiagen	DNA Mini kit (50)	51304	£169.00	50	£3.38
IDT	102 Primers F&R= 47 primers*		£205.37	350	£0.59
<b>ThermoFisher</b>	Qubit dsDNA HS High sensitivity (500) + Tubes	Q32856 + Q32856	£268.00	500	£0.54
<b>Total cost per sample (96 libraries per run)</b>					<b>£36.93</b>

\*Number of primers would vary according to the number of primers included in the PCR step.

#### 7.4.1.2 Whole genome sequencing through target enrichment

The ratio of HPV DNA to human DNA is minimal; therefore, amplification of HPV genetic material, either through PCR cycles or capture with RNA baits is required to generate enough starting material for WGS.

*7.3.1.2.1 Whole genome sequencing by PCR amplification*

Compared with genotyping by L1 sequencing, whole genome sequencing using the MiSeq requires smaller run sizes/batches due to the number of reads and coverage required. The longer the target genome, the lower the number of samples per run. For example, a WGS run with 48 samples would cost £44.73 + £5.50 of Tapestation analysis per sample sequenced (Table 33). If the number of samples included per run increased to 72, the cost per sample would reduce to £39.53 + £5.50 of the Tapestation.

**Table 33. The cost associated with NGS by PCR target enrichment (WGS).** Cost of reagents obtained in December 2022.

<b>Main reagents</b>						
<b>Supplier</b>	<b>Item</b>	<b>Reference</b>	<b>Unit Price</b>	<b>n samples</b>	<b>Cost</b>	
<b>Qiagen</b>	Qiagen Multiplex PCR Kit (100)	206143	£179.00	48	£3.73	
<b>Illumina</b>	Illumina® DNA Prep, (M) Tagmentation (96 Samples)	20018704	£1,836.29	96	£19.13	
<b>Illumina</b>	Nextera™ DNA CD Indexes (96 Indexes, 96 Samples)	20018705	£ 169.82	96	£1.77	
<b>Illumina</b>	MiSeq Reagent Kit v2 (500cycle)	MS-102-2003	£748.90	48	£15.60	
<b>Illumina</b>	MiSeq Reagent Kit v2 (500cycle)	MS-102-2003	£748.90	72	£10.40	
<b>Qiagen</b>	DNA Mini kit (50)	51304	£169.00	50	£3.38	
<b>IDT</b>	102 Primers F&R= 47 primers*		£205.37	350	£0.59	
<b>ThermoFisher</b>	Qubit dsDNA HS High sensitivity (500) + Tubes	Q32856 + Q32856	£268.00	500	£0.54	
<b>Total cost per sample (48 libraries per run)</b>						£44.73
<b>Total cost per sample (72 libraries per run)</b>						£39.53

\*Table 33 considers the number of primers designed for 1 HPV type. If more types were to be included in the detection, more primers would be required, and therefore an increase in primers costs would be expected.

#### *7.3.1.2.2 Hybridisation capture*

Hybridisation capture using designed baits is the best option for processing samples when the HPV type in the sample is unknown or when multiple HPV infections are present. Hands-on time is increased due to the hybridisation step in the last part of the library prep, as some protocols require an overnight incubation. Therefore, the time from sample to sequencing could be significantly extended compared to direct sequencing or PCR target enrichment. In contrast, PCR target enrichment requires a 3–4-hour PCR, and in some labs, these PCRs are set up to run overnight. Therefore, in practice, the use of hybridisation capture may not have such a significant impact on turnaround times as the raw numbers suggest.

The significant disadvantage of this approach concerning PCR target enrichment, is the cost of the designed baits. For this chapter, a quote was requested from a commercial provider for the library kits and bait design for up to 200 HPV types. The quote without any special discounts for 300 samples was £71,205. Sequencing 300 samples using a PCR-based enrichment approach would cost less than £10,000.

Although hybridisation-based NGS has a more complex library process, it would result in a deeper and complete analysis as it can sequence most HPV types identified (>200 types).

#### *7.3.1.3 Direct sequencing NGS*



The simplest way of detecting HPV through NGS is by performing NGS directly from all the nucleic material extracted from the specimen. This approach allows sequencing of all HPV types present in the sample and not just the ones targeted by designed primers/baits, allowing the detection of potential novel types. Moreover, direct sequencing avoids the enrichment steps, reducing enrichment biases and the sample-result time. In this case, we will save approximately 4-5 hours of the enrichment PCR, being able to perform the extraction and the library prep in an 8-hour shift.

However, this approach has some disadvantages that make it not feasible for cases where high coverage of the HPV genome is required. When considering the proportion of HPV in a tissue, the HPV nucleic acid material proportion of the samples is much lower than in humans. HPV genome is just below 8000 bp, while the human genome is 6.4 billion bp. Therefore, the small amount of HPV nucleic material will mean that number of HPV will provide low coverage depth and potentially missing regions. The low coverage would limit the WGS and whole genome analysis. Consequently, if HPV types present in the sample are already known, this is not the best approach to complete the viral genome analysis (including lineage/sub-lineage identification and integration analysis).

**Table 34. The cost associated with direct NGS (24 samples). Includes reactions for extraction, library prep and sequencing (MiSeq). Cost of reagents obtained in December 2022.**

<b>Main reagents</b>					
<b>Supplier</b>	<b>Item</b>	<b>Reference</b>	<b>Unit Price</b>	<b>n samples</b>	<b>Cost per sample</b>
Illumina	Illumina® DNA Prep, (M) Tagmentation (96 Samples)	20018704	£1,836.29	96	£19.13
Illumina	Nextera™ DNA CD Indexes (96 Indexes, 96 Samples)	20018705	£ 169.82	96	£1.77
Illumina	MiSeq Reagent Kit v2 (500cycle)	MS-102-2003	£748.90	24	£7.80
Qiagen	DNA Mini kit (50)	51304	£169.00	50	£3.38
<b>Cost per sample (24 libraries per run)</b>					<b>£55.48</b>

To increase the number of reads per sample, reducing the total number of libraries included in each sequencing run may be required. In this case, for a 24-sample MiSeq run, the approximate total cost, including extraction) the cost per sample will be £55.48 (Table 34). Furthermore, we will need to add a total of £5.50 to the total cost for the Tapestation analysis of the ladder and the library pool.

#### *7.3.1.4 Long read sequencing*

Long-read sequencing can be achieved by Oxford Nanopore and Pacific Biosciences (PacBio) technologies. However, due to the high capital required to purchase a PacBio instrument, Oxford Nanopore Technologies (ONT) has been the long-read sequencing approach of choice for many laboratories. ONT has already been used for HPV sequencing<sup>225–227</sup> to identify the HPV sequence and potential integration and integration breakpoints.

Table 35 contains the cost of the necessary reagents and the cost per sample of a long-read sequencing based on the Nanopore technology. With this approach, fewer samples can be sequenced per flow cell. When comparing Illumina direct sequencing with nanopore (24 samples), the cost is (£91.0 vs £60.3). Despite the cost differences, the advantages of this technology need to be considered. The main advantage is the sequencing in real-time, allowing sequencing to stop when the coverage required has been achieved. Additionally, long-read sequencing has a better resolution to identify integration and breakpoints<sup>228</sup>.

**Table 35. The cost associated with long read sequencing (ONT).** Includes reactions for extraction, library prep and sequencing (Nanopore) for 24 samples. Cost of reagents obtained in December 2022.

<b>Primary reagents - Long Read</b>					
<b>Supplier</b>	<b>Item</b>	<b>Reference</b>	<b>Unit Price</b>	<b>n samples</b>	
<b>Qiagen</b>	DNA Mini kit (50)	51304	£169.00	50	£ 3.38
<b>ThermoFisher</b>	Qubit dsDNA HS High sensitivity (500) + Tubes	Q32856 + Q32856	£268.00	500	£ 0.54
<b>Oxford Nanopore</b>	Flow Cell		£720.00	24	£30.00
<b>Oxford Nanopore</b>	Library Price	SQK-LSK112	£480.00	24	£20.00
<b>New England BioLabs</b>	NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing	#E7180	£890.00	24	£ 37.08
<b>Total cost per sample (24 libraries)</b>					<b>£ 91.00</b>

#### 7.3.1.5 Comparison of NGS approaches

When considering the cost of all the NGS approaches, the cheapest approach is the genotyping through target enrichment of a part of the HPV genome (Table 36). In this case, a MiSeq run can sequence up to 96 samples with a cost per sample of £36.93 (Table 36). Using RNA baits has a higher cost per sample (£237.35). However, it is the approach that can sequence the whole genome of most HPV types identified so far (~200) with full coverage (except potential integration) and high depth of reads. Currently, most NGS-developed assays have a target enrichment approach using PCR amplification. In this case, the cost per sample stands at £44.73.

**Table 36. Cost per sample for the NGS approaches described.** The number of libraries per run(n) varies depending on the maximum number of libraries that could be pooled and added to the sequencers.

<b>NGS approaches</b>	<b>Instrument</b>	<b>Cost per sample</b>
Target enrichment PCR-based L1 (n=96)	Illumina MiSeq	£36.93
Target enrichment PCR-based WGS (n=48)	Illumina MiSeq	£44.73
Direct sequencing (n=24)	Illumina MiSeq	£55.48
Long read Seq (24 samples)	Oxford Nanopore	£91
Target enrichment RNA baits-based	Illumina MiSeq	£237.35

Every NGS approach described in this chapter has advantages and disadvantages, but proper "fitness for purpose" relies on what we want to obtain from the NGS.

Table 37 describes the advantages and disadvantages of every approach. Direct sequencing is the best approach if we want to study the integration of HPV in the human genome or detect new HPV types, however, lack of enrichment could lead to low coverage. Moreover, samples would contain human reads and therefore, specific and appropriate governance checks and processes to address and manage this are required.

Target enrichment approaches solve the low coverage; however, using primers or designed probes could mean new variants or types are missed. Moreover, enrichment steps would mean no human reads in the fastq files, making it easier to share or submit to public repertoires.

**Table 37. Advantages and Disadvantages of each of the NGS approaches**

<b>NGS approaches</b>	<b>Advantages</b>	<b>Disadvantages</b>	<b>Potential Applications</b>
<b>Direct sequencing</b>	No PCR pre library required Less hands-on-time No limited by primers Can identify novel HPV types	Low proportion HPV: Human DNA Fewer samples per MiSeq run Lower coverage / missing regions?	Integration, novel types discovery
<b>Target enrichment PCR-based</b>	High coverage/depth A higher number of samples per run Deep sequencing	PCR amplification prior to library prep Design of primers Less sensitive to multiple types of infection PCR error transferred to sequencing Can only detect those types included in the primers	A deep study of already-known HPV Integration analysis ctDNA analysis R&D: biomarkers
<b>Target enrichment RNA baits-based</b>	High coverage/depth A higher number of samples per run Deep sequencing	Capture step during/after library prep Design of RNA baits Very expensive	A deep study of HPV types not known ctDNA analysis R&D: biomarkers
<b>Long read Seq</b>	Lower coverage/depth than Illumina Can be done with PCR enrichment Real-time sequencing	Higher error rate More expensive for low number of samples	Integration Quick sequencing

Long-read sequencing has progressed, and the error rate has been reduced considerably. However, it is still higher than the error rate in the short-read technology, but actual time sequencing could lead to a faster turnaround time. Moreover, it allows re-use the of the flowcell, which means smaller sample batches are required.

### 7.3.2. Comparison of conventional HPV tests and NGS

#### 7.3.2.1 Cervical screening/disease management.

Due to the large number of samples required for cervical screening/disease management, NGS is does not serve as a replacement for screening at the moment. In 2018/19, 407,854 cervical screening tests were processed in Scotland<sup>229</sup>. These days, the conventional HPV tests approved for cervical screening have high sample capacity, some offer concurrent genotyping and many incorporate pre-analytics and nucleic acid extraction. Additionally, complexity and turnaround time (days vs hours for conventional tests) make the actual NGS unsuitable for high-risk HPV cervical screening. Additionally, low- and medium-income countries need simple and cheap tests, making NGS not suitable for these countries<sup>230</sup>.

#### 7.3.2.2 Immunisation surveillance

Despite the fact that population-based immunisation surveillance requires a smaller number of samples tested than cervical screening, immunisation surveillance still tests a large number of samples. For example, when the changes in the prevalence of HPV following the vaccination in Scotland were assessed (Kavanagh *et al.*, 2017), a total of 8584 samples were genotyped<sup>76</sup>. As it occurs with the screening, NGS is unsuitable to replace genotyping with conventional HPV typing tests due to the higher cost for the same result. However, NGS can be used to sequence the HPV detected in those cases with clinically significant breakthrough infections associated with high-grade lesions and cancer even after receiving the full vaccine schedule. This approach would sequence the entire genome of the HPV (direct sequencing or hybridisation capture) to determine if the HPV, which has escaped the immune response generated by the vaccines, has any unique characteristics in its genome. Godi *et al*, found that some lineages and sub-lineages of types 33, 52 and 58 had a reduced sensitivity to monoclonal antibodies<sup>231</sup>.

#### 7.3.2.3 Annotation of cancers

As described previously, HPV-positive cancers tend to do better than HPV negative cancers. Annotation of cancers has relied on genotyping by conventional HPV tests, where an HPV-type result can be obtained quickly and at a low cost. The disadvantage of using conventional HPV tests is that these tests can only detect the specific HPV types they have been designed for. Some tests can detect the main high-risk (16, 18, 31, 33, 35, 39, 45, 52, 58 and 68), others can only detect 16 and 18 + other high-risk as a group, and some others can detect high-risk and low-risk types (6,11, 40, 42, 43, 44, 54, 61 and 70) types. The limitation is that HPV test are designed to detect only specific HPV types and therefore types not included or novel types (not discovered yet) would be missed by the test. Here is where NGS has an advantage, as direct sequencing can detect all types and identify novel types.

In this case, NGS would be used to detect those cancer cases HPV negative on conventional tests. Doing direct sequencing will ensure that no other HPV types are present, potentially causing the lesion. Additionally, by performing WGS, it would be possible to identify if there has been an integration of the original test target that could have resulted in a negative result.

#### *7.3.2.4 Deep interrogation of the HPV genome: Mixed populations and variants*

As conventional HPV tests do not provide information other than the type detected, deep interrogation of the viral genome is where NGS offers value over conventional tests. NGS can be used following different protocols but will always provide more information than conventional tests. If a target sequencing approach (e.g., focused on L1 only) is used, it will be possible to obtain detailed information on the L1 gene and, if it is complete, any SNPs or other details that a conventional test would miss. If WGS is performed, the amount of data obtained will help identify novel types or lineages and sub-lineages. Moreover, it would be possible to determine if there is an integration event and determine what genes are missing or partially missing.

#### *7.3.2.5 Research and development, including biomarker design and applications.*

Again, this application is where NGS makes a difference compared to conventional tests. Deep sequencing allows the identification of minimal variations in the viral genome (SNPs). Some SNPs have been associated with a higher risk of persistence (350T in E6<sup>51</sup>). Moreover, thanks to the high sensitivity of NGS, it can be used to quantify the viral load in samples. As published and confirmed in this thesis, viral load may play a role in prognostication. NGS can be used not only to analyse the genome of the HPV but also to quantify the number of copies present in the sample.

#### *7.3.2.6 Liquid biopsy detection (circulating HPV DNA and or circulating tumour DNA)*

One of the advantages of NGS is its high sensitivity. This makes it perfect for those diagnostic tasks where sensitivity is essential. This is the case for the detection of cell-free DNA. An increasing number of studies have found that by looking at the presence of HPV DNA and or tumour fragments in circulating blood in patients under treatment, majority of them performed in oropharyngeal cancers<sup>17–20</sup>. The absence of cell-free HPV DNA over time suggests that cancer has been eradicated and has a lower probability of relapse.

If the HPV type is known, target enrichment is the best approach to try to find the cell-free HPV DNA in the blood of the patient, thanks to the high sensitivity it can offer. Another novel test that surpasses the sensitivity of conventional tests that has been used to detect HPV ctDNA is the ddPCR. However, NGS has at least the same sensitivity and can also detect other mutations than the ddPCR cannot<sup>232,233</sup>.

#### *7.3.2.7 Summary – NGS vs conventional tests*

Even if NGS is used more frequently in diagnostics, it is not suitable for all HPV testing requirements at the moment. Due to the high cost and complex process (both wet and dry lab), it is only suitable for those cases where the extra information NGS provides is necessary/helpful. Table 38 describes the most suitable



HPV testing option for different cases. Best options were identified by looking at the approach that would provide the most information with the simplest protocol.

**Table 38. Best HPV test by objective (conventional vs NGS).**

<b>Objective</b>	<b>Most suitable option – HPV testing</b>
Screening/disease management	Commercial HPV assays
Immunisation surveillance	Commercial HPV assays/ NGS in vaccinated HPV-positive cases with clinical manifestation.
Annotation of cancers	Commercial HPV assays/ NGS in HPV-negative for confirmation
Deep interrogation	NGS – direct sequencing/enrichment RNA baits
R&D and biomarkers	NGS – direct sequencing or target enrichment if type known.
ctDNA	NGS – target enrichment due to the high sensitivity it offers.

At the moment (2023), the high cost, lengthy and complex library prep process and the extended turnaround time associated with NGS makes it not cost effective/useful for large population screening or immunisation surveillance. Genotyping by NGS provides the same data than conventional PCR tests but with a higher cost and longer turnaround times. Therefore, NGS does not seem to be a replacement for conventional HPV tests for screening or immunisation surveillance at the moment.

Where NGS is superior to conventional PCR test is in the annotation of HPV negative cancers. Direct sequencing can detect integration (potentially on the PCR target), HPV types not included in the PCR panel used or novel types. Due to the small number of HPV-negative cervical cancers, NGS could be easily implemented in a cervical screening protocol, and ensure HPV-negatives are truly negative. In those cancers with a very high positivity of HPV (cervical, anal), HPV-negative cases should be tested by NGS to entirely discard the association of the virus with the lesion. At the moment there are no guidelines recommending NGS on HPV negative cervical cancers, however, a manuscript is under preparation recommending NGS as the last resource for those HPV-negative cervical cancers. For this case, the potential presence of unknown HPV types or potential novel types make direct sequencing the method of choice.

In addition, deep interrogation of the HPV genome can only be achieved by the NGS technology. Massive parallel reads allow detecting integration of the viral genome into the host, different SNPs associated with higher persistence of infection but also the identification of biomarkers. And only a target enrichment protocol can reach the high depth of coverage required.

Moreover, the higher sensitivity of NGS technology (with respect to PCR conventional tests) allows the detection ultra low concentrations of HPV ctDNA, potentially present in blood samples obtained from patients under treatment for HPV-positive cancers. As in these cases the HPV type is already known; PCR-based target enriched NGS approach should be the method of choice.

### *7.3.3 Case study: Potential Implementation of HPV sequencing using the existing WGS infrastructure inherited from the COVID-19 pandemic.*

#### *Introduction*

Due to the COVID-19 pandemic in 2020-2022, UK and international health services/organisations and diagnostic laboratories purchased different sequencing platforms to understand, identify and investigate the different SARS-CoV-2 variants in the population. When this chapter was prepared (February 2023), the number of COVID-19 cases had reduced in the UK, and the large sequencing capacity installed over the UK nations was being underused. Laboratories and management are now looking into extending this sequencing capacity to other organisms/departments and using the capital investment performed in sequencing equipment and associated systems.

This case scenario will describe a direct sequencing approach (broad spectrum) to perform WGS of HPV. This approach was selected from the approaches described above due to its simplicity (compared to other approaches) and the multiple applications. The case scenario will include a description of sample types and sampling, laboratory requirements, sequencing process, and bioinformatic analysis by describing the

instruments most commonly available in a diagnostic laboratory as a consequence of COVID-19 as an example.

### *Sample type and sampling*

HPV sequencing can mainly be performed using two different types of samples, biopsy sections and liquid-based cytology (LBC) samples. However, other specimens have resulted valid for HPV sequencing, such as circulating blood and swabs<sup>220</sup>.

The first step of the HPV sequencing process is obtaining the nucleic acid. From LBC, nucleic acid can be extracted using conventional or automated methods (e.g., Qiagen, Biomerieux or Seegene). However, biopsy sections tend to be preserved as formalin-fixed paraffin-embedded (FFPE). Due to the nature of FFPE, most automated extraction platforms are incompatible. Paraffin, used to preserve the tissue, can affect the instrument by clogging the tubes/pipette system. Thus, a pre-extraction treatment<sup>191</sup> prior to the automated or manual extraction approach is the best option. Prior to starting the library prep, it is recommended to check the quality, quantity and size of the DNA. A fragmented and degraded DNA could result in a sub-optimal library prep leading to a low DNA yield and low cluster density on the sequencing instrument.

### *Library prep – direct sequencing*

As we are performing broad-spectrum sequencing, there is no need to target enrich HPV DNA prior to the sequencing library. In library prep, DNA gets fragmented and barcoded so the sequencing instruments can read it. It usually involves long hand-on-time (it varies depending on the kit used), taking from 1-2 to 7-8 hours. Multiple companies offer different library kits; however, only a few are compatible with liquid-handling robots that can perform library preparation.

These days there are autonomous liquid handling robots that can perform the library prep process, like the Hamilton NGS STAR robot (Reno, USA). Use of these robots results in a higher number of libraries performed per day, reduced time, and reduced potential pipetting and human errors. Access to one of these robots could be very advantageous; however not mandatory to obtain good quality libraries.

#### *Sequencing process/instruments*

Nowadays, there are multiple sequencing instruments available. For example, Illumina has different instruments in its catalogue, with different characteristics (number of reads and amount of data produced). They mainly vary in the amount of data they can produce, which relies on the number of libraries (number of samples) they can have as input. A NextSeq 550 instrument run can include a maximum of 384 different libraries (microorganism sequencing), while for a MiSeq, it is limited to 96. For the approach of choice, 24 samples is the maximum number of samples per run, with 16 samples potentially giving the best coverage necessary to perform a deep analysis of HPV. The number of samples depends on the number of copies of HPV present in the sample.

#### *Bioinformatic analysis*

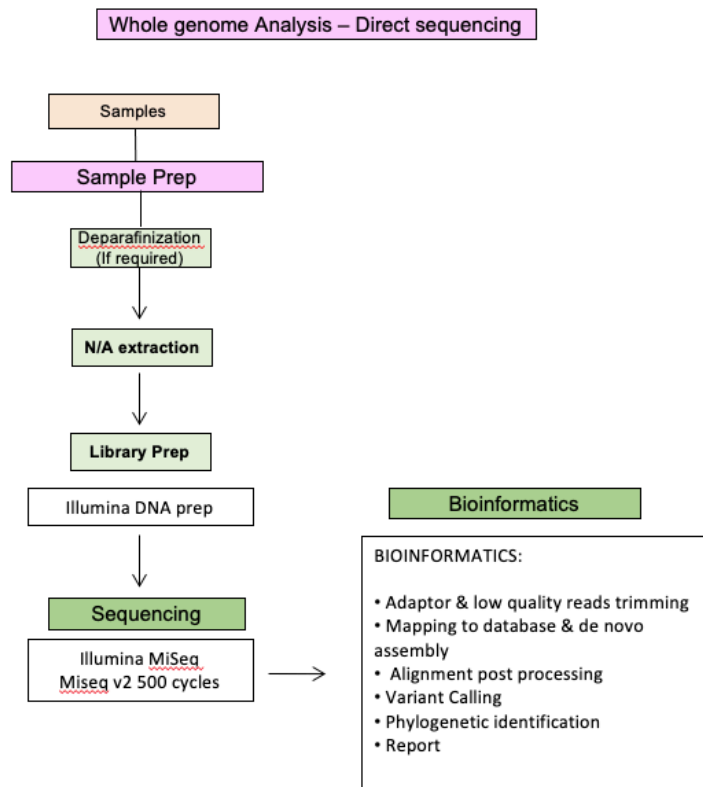
As with other molecular tests, WGS does not end with the sequencing step; it requires data analysis using various bioinformatic tools. Typically, the analysis demands powerful CPUs/processing capacity that most consumer computers do not have. Departments or institutions can access cloud servers where analysis and calculations are performed or have an in-house computer/GPU with enough processing power to perform the analysis.

Regarding the bioinformatic tools/pipelines, the majority of the tools are available for free on the internet. Nevertheless, advanced computer and bioinformatic skills are required to install and use these tools.

Having a bioinformatician, or bioinformatics department support would be very beneficial for implementing and maintaining the analysis pipeline.

A bioinformatic pipeline for HPV analysis must include the following steps (Figure 52):

- Pre-processing and QC
- Reference or *de novo* assembly
- Reference HPV types database
- Variant and mutation analysis
- Lineage assignment (optional)



**Figure 52. Description of the necessary steps to perform a direct sequencing approach for HPV investigation.**

As an example, a pipeline for direct sequencing and detection of multiple HPV types would include the following:

- Quality: Reads check and adaptor trimming by Trimmomatic
- Database for alignment (database containing fasta files of all the HPV types discovered)
- Alignment to reference: Alignment using an aligner tool using the in-house HPV database as the reference.
- de novo assembly
- Filtering human reads: Samtools or NextGenMap.
- Check the depth of coverage. For type identification, >20-40x would be valid.
- Variant calling: detection of SNPS and indel. GATK HaplotypeCaller

#### **7.4 Discussion**

The decision to choose the best NGS approach for HPV diagnosis relies on the necessities of the specialist laboratory and the type of information required. This chapter has presented the different NGS approaches, applications for HPV sequencing and how they compare with standard PCR based HPV tests.

Even though NGS cost has reduced over the years with promises of sequenced genomes for \$100, it is still an expensive technology. Additionally, complex and lengthy protocols and longer turnaround times make NGS not a valid option to replace conventional genotyping HPV tests wholesale. However, NGS can be used to complement the diagnosis, obtaining information from the HPV that conventional tests cannot obtain. This includes the detection of mutations associated with higher infection persistence, identifying sub-lineages with a higher risk of cancer, novel types, novel mutations and integration of the virus in the human genome (associated with worse survival outcomes). In addition, higher sensitivity than conventional tests makes NGS a handy tool for detecting low levels of HPV infection or circulating DNA in the blood.

From a diagnosis point of view, there are two NGS protocols that can be used in a HPV diagnosis lab at the moment. One is the use of direct sequencing. It can be used to detect any HPV or confirm the absence in previously diagnosed HPV-negative cancers due to the capacity of detecting multiple and novel HPV types (or not included in the genotyping test used previously). Target enrichment must be the protocol of choice when a deep analysis of the virus is required or for the detection of ctDNA in patients under treatment for HPV positive cancers and determine the efficacy of the treatment.

In addition to the sequencing, we must remember the bioinformatic analysis, an essential part of the sequencing, where we process and interpret the data generated. Most of the tools required to do the analysis are open-source and available for free from online repositories. However, bioinformatics and command line knowledge are required to install and use most of these tools. Access to a bioinformatics service/department would simplify the process and maintenance of pipelines.

Also, due process of governance and ethical considerations and approvals must be obtained for data access and transfer to external servers due to human genome reads (if a direct sequencing approach is followed).

As NGS is still new in the HPV diagnosis field, there are no international guidelines or recommendations for the use of NGS in HPV diagnosis laboratories. Due to this lack of information, a paper was prepared between the Scottish HPV reference laboratory and the HPV International reference center, describing the importance of validation and quality assurance for WGS NGS HPV sequencing<sup>220</sup>. At the time this chapter is prepared (April 2023), it has been proposed (manuscript under preparation) that those HPV-negative cervical cancers (using standard >1 PCR approach) should be triaged to NGS. This is the first acknowledged NGS application by the HPV community.

The case scenario presented one of the approaches that could be implemented in an HPV diagnostic laboratory. Direct sequencing was the selected approach thanks to the multiple uses (and advantages with respect to conventional PCR based tests), from detecting the broad range of HPV types but also new types, to identify integration parts of the HPV genome in the host.

All the data generated by a NGS protocol has different potential end users. From the clinical scientist that would identify the types/sub-lineages or integration to the clinician that would evaluate the presence of specific biomarkers in the HPV and consider a specific treatment. But also, to Public Health where data could help identifying HPV changes in the population (due to the immunisation pressure) as well as performing vaccine surveillance.

Little by little, NGS is becoming increasingly prevalent as a standard method for diagnostic genotyping in national reference laboratories. It can provide a large amount of information and impact the management of patients suffering from HPV disease. However, there are still labs that do not have implemented it yet, one of the reasons being the complexity of the process. But simply by sharing protocols and experiences between laboratories, it would be easier to laboratories to implement this powerful technology. The more HPV genome data is generated, the greater ability to investigate the clinical correlations and vulnerability of cancers caused by HPV that could eventually lead into enhance patient protection and improve outcomes for HPV-induced cancers.



## 8. Final Discussion

### *Summary of findings*

Despite the great effort in place in Scotland to reduce and prevent HPV-driven cancers through immunisation, screening and research, there were still some gaps in anal cancer and biomarkers identified that this thesis has tried to address. In addition, due to the increased use of NGS, an overarching aim of this thesis was to undertake research demonstrate the utility of genomics for providing and new knowledge and insight into the HPV disease in Scotland and identify the potential applications and use NGS for HPV analysis for reference services.

This study explored the type-specific diversity and prevalence of HPV in Scotland's most common HPV-driven cancers: cervical, oropharynx and anal. For cervical cancers, it was identified that almost 92% were positive for HPV, 94.4% for the SCC group and 83.61% for ASC+ADC. In the oropharyngeal cancers collected between 2013 and 2020, HPV prevalence identified was 55%. High-risk types were identified in 54% of the cases. This prevalence aligns with the ones obtained in other national studies. Schache *et al*<sup>109, 1</sup> identified that the overall proportion of HPV positive OPSCC between 2002 and 2011 in the UK was 52%. As expected, HPV 16 was the dominant type in the Scottish data, detected in 51% of the OPC tested and 93% of the total positives. Despite increased HPV-associated oropharyngeal cancers in the USA<sup>134</sup>, no rise has been identified in Scotland.

Due to the lack of information about HPV prevalence in anal lesions in Scotland, this thesis looked at the HPV prevalence in a population-based cohort of anal cancer collected over ten years. Most anal cancer samples were positive for at least one HPV type (89%) and HPV 16 was the dominant type (83%) in anal cancer-positive cases, agreeing with the high positivity of HPV in anal cancer and the high prevalence of

HPV 16 reported by De Sanjosè *et al.*<sup>131</sup>. Demographic analysis of HPV in anal cancers showed a significant association between HPV-positivity and females, with HPV-positive prevalence in women higher than in males (66.46% vs 33.53%). Data also showed that HPV-positive status is highly associated with those alive at the time of data censoring in anal cancer cases. By looking at overall survival, HPV-positivity was associated with improved overall survival. This aligns with the systematic review by Urbute *et al.*, where they found HPV-positive anal cancer had significantly better overall survival than HPV-negative<sup>161</sup>. This is similar to an emerging pattern in other cancers associated with HPV, including cervical<sup>173,174</sup>, oropharyngeal<sup>175,176</sup>, penile<sup>177,178</sup> and vulval cancers<sup>179</sup>.

The viral load of HPV 16 in anal cancers was obtained by using a ddPCR approach and used to analyse any association with overall survival. However, adjusted Cox HR showed that viral load did not influence survival. However, as the confidence interval was just above 1, it could be plausible that the association could tip into significance by performing a more extensive study.

Prior to starting with the NGS and WGS, three different nucleic acid extraction methods were assessed to determine which one could be better for downstream NGS. The Qiagen FFPE kit obtained the higher DNA concentration and the best result on the qPCR comparison. However, due to the NGS approach selected (target enrichment) differences could be reduced by the amplification of the DNA prior to the library prep. If a different NGS approach would be followed, the QIAamp FFPE tissue kit could obtain the concentration and fragment size required for short-read sequencing.

Whole genome sequencing was performed on those anal cancer samples positive for HPV 16. In addition, HPV 16-positive residual rectal swabs from a MSM population were used as a "control" group. A1 and A2 sub-lineages were the most dominant in both cohorts. Only minimum differences in lineages B and C were

identified. The lack of sub-lineage diversity and absence of those sub-lineages associated with a higher risk of cancer could be linked to the demographic factors and the composition of the study cohort, such as the ethnicity that where the population in Scotland is predominantly white. Furthermore, no significant differences in overall survival were identified for A1 and non-A1 sub-lineages. This suggest that HPV 16 sub-lineage identification may not be a useful biomarker (at least in Scotland). However viral load data looks encouraging and could potentially be used as a biomarker to identify those cases that could require a more aggressive treatment or a more regular follow-up. Use of biomarkers would not only apply to anal lesions, but also in cervical and oropharyngeal lesions.

The last part of the thesis presented the different applications of NGS and how they compare with conventional HPV tests. Besides, a case study scenario for the application of WGS in has been presented describing the different parts that need to be considered when implementing NGS (direct sequencing approach) in a specialist laboratory. Due to the high cost associated with NGS, lengthy process and extended turnaround times, conventional genotyping is still the most convenient approach when only the qualitative or quantitative detection of the most common HPV types is necessary (like cervical screening or immunisation surveillance). However, in those cases where the deep analysis of the viral genome is required, NGS and more specifically a target enrichment protocol overpasses a conventional PCR tests. In addition, NGS direct sequencing is the best approach should be the approach selected for the testing of HPV-negative lesions.

#### *Implications of the work - When will we see the impact of vaccination?*

HPV-type prevalence data provides information on the proportion of cases that could be prevented through vaccination and gives insight into the extent of disease that is unlikely to be reduced. Currently, the HPV vaccine offered in Scotland and the United Kingdom is Gardasil 9 (started in the 2021/22 school

year). However, the bivalent and quadrivalent vaccines were offered from 2008 to 2012 and 2012 to 2021/22, respectively. When considering the types present in cervical cancer lesions, any of the three vaccines could potentially prevent 86-87% of cases when factoring in cross-protection). For the OPC cohort, 52-54% of HPV-positive cases could be prevented with any of the three vaccines. The new anal HPV prevalence suggests that 85-88% of anal cancers in Scotland could also be prevented.

Data collated in this thesis has shown that age plays a role in HPV infection. For those cervical cancer samples where age information was available, it has been identified that HPV positivity declines with age, detecting the highest HPV positivity in women aged <45 (96%), decreasing in the older groups to 81% in women aged ≥75 years. In OPC, data showed that the prevalence of HPV-positive cases tends to be higher in those <60 years old than in the older population. Overall HPV positivity was 69.59% in those <50 years old, decreasing to 50% in 60-69 and 35% in 70<sup>3</sup>. For the anal cancer cohort, most anal cancer cases have been diagnosed in >50 years old<sup>84,171</sup> (85%), and almost 60% in > 60 years old. The reasons for the reduction of HPV prevalence with age and the increase in non-HPV cervical cancers are not fully understood. It is plausible that there may be a more significant opportunity for the HPV to be “lost” during the carcinogenic process in the elderly<sup>112</sup>, or maybe with ageing, there is a greater chance that non-HPV cancerous changes/pathways may play a larger role<sup>113</sup>. In OPC, we know non-HPV OPC increases with age likely to have drunk and smoked heavily, both critical risk factors<sup>114</sup>, more common in older men, with a median age of 61<sup>115</sup>.

For cervical cancers, the effect of the vaccines has already been detected in England<sup>234</sup> and Sweden<sup>235</sup>. The English study found a relative reduction in cervical cancer rates for those vaccinated at age 12-13 of 87% compared to the unvaccinated cohort. On the other hand, the Swedish study found that the risk of cervical cancer among those who received the HPV vaccine before the age of 17 had 88% lower risk than

among those who had never been vaccinated. Nevertheless, due to the higher incidence of HPV-associated OPC and anal cancers later in life, we will not see the full effect of the vaccines for other HPV-driven cancers for another one or two decades in women and probably longer for the male population.

#### *Implications of the work - HPV negatives*

Regarding non-HPV cases, 8.42% of the cervical cases did not test positive for any HPV using PCR-based tests. A recent publication from Arroyo Mühr *et al.*, 2020<sup>62</sup> showed that 43% of negative cervical cancers (by PCR) were positive after using NGS and that most were positive for high-risk or probably high-risk HPV types. This suggests it is feasible that some types have not been detected due to the molecular assay choice/sensitivity or possible partial or complete missing of the target region due to integration.

In this thesis, I did not have the opportunity to sequence the negative anal cancer lesions to discard any false negative case. However, I have identified that, as seen in other HPV-driven cancers, HPV negative anal cancers, have a worse overall survival than those HPV positive cases. There are other HPV-driven cancers, where it also needs to be confirmed, but HPV status seems to play an important role in the prognosis.

Thus, the realisation of the importance of a correct diagnosis of those HPV-negative cancers in the HPV community has led the main national HPV reference laboratories to prepare a communal protocol for those HPV negative cancers. Recently, a paper has been prepared in collaboration of some of the leading HPV laboratories in the world proposing international guidelines to standardise the identification of HPV negative high grade squamous intraepithelial lesions and standardise the re-analysis to discard HPV presence in the lesion: Petret *et al.*, (2023) Human Papillomavirus negative high-grade cervical lesions: A

suggested guideline for HPV testing quality assurance). These guidelines, suggest different testing on HPV negative samples, including WGS, to completely rule out the presence of HPV.

#### *Implication of the work – NGS & WGS*

Similarly, while I was learning about NGS and HPV, looking for guidelines I noticed a gap in the literature, with very little information publicly available. Therefore, in collaboration with the Karolinska institute, we published a paper where we described recommendations for validation and quality assurance procedures of each step of a NGS workflow, with a focus on WGS approaches.

In addition, the case study presented in the last chapter describing a direct sequencing approach could hopefully be helpful for HPV specialist labs that are trying to implement a broad sequencing approach. Using the experience acquired from the HPV 16 WGS, the Scottish HPV Reference lab will implement a broad range sequencing test, where all HPV types can be detected, aiming to interrogate HPV negative cancers.

Despite the non-significant association of HPV 16 sub-lineages and overall survival, WGS has allowed the identification of samples with potential integration of HPV in the host genome, which has been associated with worse outcomes.

#### *Challenges/Limitations – What would I have changed?*

By analysing the thesis once completed, different limitations have been identified. They range from missing data to funding available to time. In the cervical cancer data, a large number of samples did not contain histology or age information. ISD was contacted to retrieve the missing information from their databases, but by the time this chapter was prepared, the information had yet to be received. COVID-19 has also impacted ISD, and resources were focused on COVID-19 data analysis. Therefore, the analysis of

age and histology needed to be completed due to the missing data. Another limitation identified is the presence of only male samples in the control cohort used in the sub-lineage comparison. These samples were the only ones available; however, for future investigations, it would be advantageous to have a combination of both sex for a better representation.

Another main limitation of this thesis has been the high cost of the NGS reagents and the funding available to conduct research for this thesis. Although NGS only focused on anal cancers and HPV 16 cases, it would have been fascinating to sequence from HPV-negative cervical cancers to discount the presence of the HPV thoroughly. Similarly, testing those HPV present in ASC+ADC cases, and determine if they do have any difference from those types present in SCC cases would have potentially been informative. In addition, it would have been fascinating to determine the sub-lineage of HPV 16 of the OPC and determine if there are driven only by one of the multiple sub-lineages, and also for evidence of integration analysis of OPC and CCX.

However, with the limitations on research funding, samples available, bioinformatic analysis resources and time, especially in with the impact of COVID-19 on the ability to conduct laboratory experiments, choices had to be made and investigations were targeted accordingly.

HPV testing is not routinely performed on anal lesions in Scotland; therefore, there was no national data we could obtain. Moreover, I only tested anal lesions collected in the east region of Scotland, including Lothian, Fife and Borders. HPV information from the whole country or at least the largest region of Scotland (Greater Glasgow and Clyde) could better represent the HPV prevalence in Scotland.

In terms of changes to my PhD and with the benefit of seeing the work in its entirety, I would have performed direct sequencing in those HPV-negative anal cancers to determine if they were true negatives or if the HPV was hidden. Also, with more planning, I would have repeated the samples with missing regions with a direct sequencing approach and looked in detail at the integration of HPV in the human genome, trying to understand better the integrations sites.

#### *Future work*

As this work has only covered CCX, OPC and anal cancer, it would be interesting to analyse the HPV prevalence in other HPV-driven cancers such as vaginal, vulvar or penile. This will help understand the proportion of lesions driven by HPV, what types are driving these lesions, if changes in prevalence have registered over time, and the potential effect of the HPV vaccines. As the incidence of these cancers is low (at least in Scotland), it would be easy to capture cases from the last 5-10 years and test them for HPV. In addition, NGS analysis on samples other than anal cancers could result very helpful to understand integration.

As mentioned above, HPV 16 sub-lineage identification does not seem to be a good biomarker in Scotland. However, different studies already published, have shown that HPV status and viral load have an impact on the overall survival of the patient. Maybe future update of the diagnosis guidelines (FIGO<sup>44</sup> and TNM classification<sup>236</sup>) could add the viral load or HPV status as variables to be considering when performing the diagnosis of cancer.

Finally, the data obtained by using NGS clearly shows the power of this technology and how much we can learn from it. Analysing in deep the HPV or integration could improve the diagnosis we offer and potentially significantly impact the diagnosis or management of the patient. In a future, where majority of the population would be vaccinated against HPV, a low prevalence of high-risk HPV could make genotyping irrelevant while NGS could become the tool necessary to determine risk of cancer by



integration identification or viral load. However, further studies are required. In the meantime, deep analysis of those “HPV-negative” cases (majority in the elderly population) are required as these are cases that from individual that have not benefited from the HPV vaccines or vaccinated population and from an HPV-based cervical cancer screening program.

As we go forward and NGS becomes more widely adopted as a routine diagnostic genotyping technology, and HPV genome data becomes more widely available and linked to patient data, we will be in a better position to probe the clinical associations and susceptibility of HPV driven cancers. The combination of HPV, patient and cancer genomics and digital health data will provide valuable resources and powerful information to help protect patients and improve outcomes against cancer caused by HPV in the years to come.

## 9. References

1. Human papillomaviruses. *IARC Monogr Eval Carcinog Risks Hum.* 2007;90:371-381. Accessed September 7, 2021. <https://pubmed.ncbi.nlm.nih.gov/18354839/>
2. De Villiers EM, Fauquet C, Broker TR, Bernard HU, Zur Hausen H. Classification of papillomaviruses. *Virology.* 2004;324(1):17-27. doi:10.1016/j.virol.2004.03.033
3. Zur Hausen H. Human papillomaviruses and their possible role in squamous cell carcinomas. *Curr Top Microbiol Immunol.* 1977;78:1-30. doi:10.1007/978-3-642-66800-5\_1
4. STI Incidence, Prevalence, Cost Estimates | CDC. Accessed April 24, 2023. <https://www.cdc.gov/nchhstp/newsroom/2021/2018-STI-incidence-prevalence-estimates.html>
5. Bernard HU, Burk RD, Chen Z, van Doorslaer K, Zur Hausen H, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology.* 2010;401(1):70-79. doi:10.1016/j.virol.2010.02.002
6. Muñoz N, Bosch FX, De Sanjosé S, et al. Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *New England Journal of Medicine.* 2003;348(6):518-527. doi:10.1056/nejmoa021641
7. Burd EM. Human papillomavirus and cervical cancer. *Clin Microbiol Rev.* 2003;16(1):1-17. doi:10.1128/CMR.16.1.1-17.2003
8. Mehanna H, Jones TM, Gregoire V, Ang KK. Oropharyngeal carcinoma related to human papillomavirus. *BMJ (Online).* 2010;340(7752):879. doi:10.1136/bmj.c1439
9. Hans-Ulrich B, Burk RD, Chen Z, van Doorslaer K, zur Hausen H, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology.* 2010;401(1):70-79. doi:10.1016/J.VIROL.2010.02.002
10. Bucchi D, Stracci F, Buonora N, Masanotti G. Human papillomavirus and gastrointestinal cancer: A review. *World J Gastroenterol.* 2016;22(33):7415-7430. doi:10.3748/wjg.v22.i33.7415
11. Castellsagué X, Paavonen J, Jaisamrarn U, et al. Risk of first cervical HPV infection and pre-cancerous lesions after onset of sexual activity: Analysis of women in the control arm of the randomized, controlled PATRICIA trial. *BMC Infect Dis.* 2014;14(1):1-12. doi:10.1186/s12879-014-0551-y
12. Doorbar J, Quint W, Banks L, et al. The biology and life-cycle of human papillomaviruses. *Vaccine.* 2012;30(SUPPL.5):F55-F70. doi:10.1016/j.vaccine.2012.06.083
13. de Villiers EM, Fauquet C, Broker TR, Hans-Ulrich B, zur Hausen H. Classification of papillomaviruses. *Virology.* 2004;324(1):17-27. doi:10.1016/J.VIROL.2004.03.033
14. Burk RD, Chen Z, Van Doorslaer K. Human papillomaviruses: Genetic basis of carcinogenicity. *Public Health Genomics.* 2009;12(5-6):281-290. doi:10.1159/000214919
15. Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. *Virology.* 2013;445(1-2):232-243. doi:10.1016/j.virol.2013.07.018
16. Jeon S, Lambert PF. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc Natl Acad Sci U S A.* 1995;92(5):1654. doi:10.1073/PNAS.92.5.1654

17. Aydin I, Weber S, Snijder B, et al. Large Scale RNAi Reveals the Requirement of Nuclear Envelope Breakdown for Nuclear Import of Human Papillomaviruses. *PLoS Pathog.* 2014;10(5). doi:10.1371/journal.ppat.1004162
18. Zhang W, Kazakov T, Popa A, DiMaio D. Vesicular trafficking of incoming human papillomavirus 16 to the Golgi apparatus and endoplasmic reticulum requires  $\gamma$ -Secretase Activity. *mBio.* 2014;5(5). doi:10.1128/mBio.01777-14
19. Parish JL, Bean AM, Park RB, Androphy EJ. ChR1 Is Required for Loading Papillomavirus E2 onto Mitotic Chromosomes and Viral Genome Maintenance. *Mol Cell.* 2006;24(6):867-876. doi:10.1016/j.molcel.2006.11.005
20. Pyeon D, Pearce SM, Lank SM, Ahlquist P, Lambert PF. Establishment of human papillomavirus infection requires cell cycle progression. *PLoS Pathog.* 2009;5(2). doi:10.1371/journal.ppat.1000318
21. Graham S V. The human papillomavirus replication cycle, and its links to cancer progression: A comprehensive review. *Clin Sci.* 2017;131(17):2201-2221. doi:10.1042/CS20160786
22. Westrich JA, Warren CJ, Pyeon D. Evasion of host immune defenses by human papillomavirus. *Virus Res.* 2017;231:21-33. doi:10.1016/j.virusres.2016.11.023
23. Chellappan S, Kraus VB, Kroger B, et al. Adeno virus E1A, simian virus 40 tumor antigen, and human papillomavirus E7 protein share the capacity to disrupt the interaction between transcription factor E2F and the retinoblastoma gene product. *Proc Natl Acad Sci U S A.* 1992;89(10):4549-4553. doi:10.1073/pnas.89.10.4549
24. Doorbar J. Molecular biology of human papillomavirus infection and cervical cancer. *Clin Sci.* 2006;110(5):525-541. doi:10.1042/CS20050369
25. Buitrago-Perez A, Garaulet G, Vazquez-Carballo A, Paramio J, Garcia-Escudero R. Molecular Signature of HPV-Induced Carcinogenesis: pRb, p53 and Gene Expression Profiling. *Curr Genomics.* 2009;10(1):26-34. doi:10.2174/138920209787581235
26. Stanley MA, Sterling JC. Host responses to infection with human papillomavirus. *Current Problems in Dermatology (Switzerland).* 2014;45:58-74. doi:10.1159/000355964
27. Amador-Molina A, Hernández-Valencia JF, Lamoyi E, Contreras-Paredes A, Lizano M. Role of innate immunity against human papillomavirus (HPV) infections and effect of adjuvants in promoting specific immune response. *Viruses.* 2013;5(11):2624-2642. doi:10.3390/v5112624
28. de Sanjose, Quint WG, Alemany L, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* 2010;11(11):1048-1056. doi:10.1016/S1470-2045(10)70230-8
29. Chelimo C, Wouldes TA, Cameron LD, Elwood JM. Risk factors for and prevention of human papillomaviruses (HPV), genital warts and cervical cancer. *Journal of Infection.* 2013;66(3):207-217. doi:10.1016/j.jinf.2012.10.024
30. Giuliano AR, Harris R, Sedjo RL, et al. Incidence, prevalence, and clearance of type-specific human papillomavirus infections: The Young Women's Health Study. *Journal of Infectious Diseases.* 2002;186(4):462-469. doi:10.1086/341782
31. Mirabello L, Yeager M, Cullen M, et al. HPV16 Sublineage Associations with Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J Natl Cancer Inst.* 2016;108(9):1-9. doi:10.1093/jnci/djw100

32. Cullen M, Boland JF, Schiffman M, et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Research*. 2015;1:3-11. doi:10.1016/j.pvr.2015.05.004
33. Clifford GM, Tenet V, Georges D, et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Research*. 2019;7(November 2018):67-74. doi:10.1016/j.pvr.2019.02.001
34. van der Weele P, Meijer CJLM, King AJ. Whole-Genome Sequencing and Variant Analysis of Human Papillomavirus 16 Infections. *J Virol*. 2017;91(19):JVI.00844-17. doi:10.1128/jvi.00844-17
35. Van Der Weele P, Meijer CJLM, King AJ. High whole-genome sequence diversity of human papillomavirus type 18 isolates. *Viruses*. 2018;10(2). doi:10.3390/v10020068
36. Nicolás-Párraga S, Alemany L, De Sanjosé S, Bosch FX, Bravo IG. Differential HPV16 variant distribution in squamous cell carcinoma, adenocarcinoma and adenosquamous cell carcinoma. *Int J Cancer*. 2017;140(9):2092-2100. doi:10.1002/ijc.30636
37. Cornet I, Gheit T, Clifford GM, et al. Human papillomavirus type 16 E6 variants in France and risk of viral persistence. *Infect Agent Cancer*. 2013;8(1):4. doi:10.1186/1750-9378-8-4
38. Cornet I, Gheit T, Iannacone MR, et al. HPV16 genetic variation and the development of cervical cancer worldwide. *Br J Cancer*. 2013;108(1):240-244. doi:10.1038/bjc.2012.508
39. Mirabello L, Yeager M, Yu K, et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell*. 2017;170(6):1164-1174.e6. doi:10.1016/j.cell.2017.08.001
40. Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR, Stanley MA. The biology and life-cycle of human papillomaviruses. *Vaccine*. 2012;30 Suppl 5(SUPPL.5). doi:10.1016/J.VACCINE.2012.06.083
41. Kreimer AR, Chaturvedi AK. HPV-associated oropharyngeal cancers - Are they preventable? *Cancer Prevention Research*. 2011;4(9):1346-1349. doi:10.1158/1940-6207.CAPR-11-0379
42. Chan JK, Monk BJ, Brewer C, et al. HPV infection and number of lifetime sexual partners are strong predictors for "natural" regression of CIN 2 and 3. *Br J Cancer*. 2003;89(6):1062-1066. doi:10.1038/sj.bjc.6601196
43. PDQ Adult Treatment Editorial Board. Cervical Cancer Treatment: Patient Version. *PDQ Cancer Information Summaries*. Published online 2002:1-22. Accessed September 7, 2021. <http://www.ncbi.nlm.nih.gov/pubmed/26389422>
44. Bhatla N, Berek JS, Cuello Fredes M, et al. Revised FIGO staging for carcinoma of the cervix uteri. *International Journal of Gynecology and Obstetrics*. 2019;145(1):129-135. doi:10.1002/ijgo.12749
45. AJCC. AJCC Cancer Staging Manual 8th Edition. *Definitions*. Published online 2020:489-539. doi:10.32388/b30ldk
46. Table of Contents page: *Annals of Oncology*. Accessed September 7, 2021. [https://www.annalsofoncology.org/issue/S0923-7534\(12\)X4900-7](https://www.annalsofoncology.org/issue/S0923-7534(12)X4900-7)
47. Glynne-Jones R, Nilsson PJ, Aschele C, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. *Radiother Oncol*. 2014;111(3):330-339. doi:10.1016/J.RADONC.2014.04.013
48. Table of Contents page: *Annals of Oncology*. Accessed September 9, 2021. [https://www.annalsofoncology.org/issue/S0923-7534\(12\)X4900-7](https://www.annalsofoncology.org/issue/S0923-7534(12)X4900-7)

49. Tarazi R, Nelson RL. Anal adenocarcinoma: A comprehensive review. *Semin Surg Oncol*. 1994;10(3):235-240. doi:10.1002/ssu.2980100312
50. Arbyn M, Simon M, Peeters E, et al. 2020 List of Human Papillomavirus Assays Suitable for Primary Cervical Cancer Screening. *Clinical Microbiology and Infection*. 2021;27(8):1083-1095. doi:10.1016/j.cmi.2021.04.031
51. Grodzki M, Besson G, Clavel C, et al. Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350g variant. *Cancer Epidemiology Biomarkers and Prevention*. 2006;15(4):820-822. doi:10.1158/1055-9965.EPI-05-0864
52. Arroyo Mühr LS, Lagheden C, Hultin E, et al. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study. *Br J Cancer*. 2018;119(9):1163-1168. doi:10.1038/s41416-018-0311-7
53. Arroyo Mühr LS, Smelov V, Bzhalava D, Eklund C, Hultin E, Dillner J. Next generation sequencing for human papillomavirus genotyping. *Journal of Clinical Virology*. 2013;58(2):437-442. doi:10.1016/j.jcv.2013.07.013
54. Yi X, Zou J, Xu J, et al. Development and validation of a new HPV genotyping assay based on next-generation sequencing. *Am J Clin Pathol*. 2014;141(6):796-804. doi:10.1309/AJCP9P2KJSXEKCB
55. Goswami K, Clarkson S, Phillips CD, et al. An Enhanced Understanding of Culture-Negative Periprosthetic Joint Infection with Next-Generation Sequencing: A Multicenter Study. *Journal of Bone and Joint Surgery*. 2022;104(17):1523-1529. doi:10.2106/JBJS.21.01061
56. Butt S, Allison L, Vishram B, et al. Epidemiological investigations identified an outbreak of Shiga toxin-producing Escherichia coli serotype O26:H11 associated with pre-packed sandwiches. *Epidemiol Infect*. 2021;149. doi:10.1017/S0950268821001576
57. Nickbakhsh S, Hughes J, Christofidis N, et al. Genomic epidemiology of SARS-CoV-2 in a university outbreak setting and implications for public health planning. *Sci Rep*. 2022;12(1). doi:10.1038/s41598-022-15661-1
58. da Silva Filipe A, Shepherd JG, Williams T, et al. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol*. 2021;6(1):112-122. doi:10.1038/s41564-020-00838-z
59. Eales O, Page AJ, de Oliveira Martins L, et al. SARS-CoV-2 lineage dynamics in England from September to November 2021: high diversity of Delta sub-lineages and increased transmissibility of AY.4.2. *BMC Infect Dis*. 2022;22(1). doi:10.1186/s12879-022-07628-4
60. Sastre-Garau X, Diop M, Martin F, et al. A NGS-based blood test for the diagnosis of invasive HPV-associated carcinomas with extensive viral genomic characterization. *Clinical Cancer Research*. 2021;27(19):5307-5316. doi:10.1158/1078-0432.CCR-21-0293
61. Arroyo Mühr LS, Lagheden C, Lei J, et al. Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR. *Br J Cancer*. 2020;123(12):1790-1795. doi:10.1038/s41416-020-01111-0
62. Arroyo Mühr LS, Lagheden C, Eklund C, et al. Sequencing detects human papillomavirus in some apparently HPV-negative invasive cervical cancers. *Journal of General Virology*. 2020;101(3):265-270. doi:10.1099/jgv.0.001374
63. de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer*. 2017;141(4):664-670. doi:10.1002/ijc.30716

64. Information Services Division A National Statistics Publication for Scotland Cancer Incidence and Prevalence in Scotland (to Information Services Division.; 2017. Accessed September 11, 2021. <https://www.statisticsauthority.gov.uk/national-statistician/types-of-official-statistics/>
65. Cameron RL, Cuschieri K, Pollock KG. Baseline HPV prevalence in rectal swabs from men attending a sexual health clinic in Scotland: Assessing the potential impact of a selective HPV vaccination programme for men who have sex with men. *Sex Transm Infect.* 2020;96(1):55-57. doi:10.1136/sextrans-2018-053668
66. Ukhsa. *Chapter 18a-Human Papillomavirus (HPV) Human Papillomavirus (HPV) Human Papillomavirus (HPV).*
67. NHS Scotland, National Statistics. HPV Immunisation Statistics Scotland: School Year 2018/2019. Published online 2019:26. Accessed September 7, 2021. <https://www.isdscotland.org/Health-Topics/Child-Health/Publications/2019-11-26/2019-11-26-HPV-Report.pdf?10793703795>
68. Cuschieri K, Brewster DH, Williams ARW, et al. Distribution of HPV types associated with cervical cancers in Scotland and implications for the impact of HPV vaccines. *Br J Cancer.* 2010;102(5):930-932. doi:10.1038/sj.bjc.6605556
69. O’Leary MC, Sinka K, Robertson C, et al. HPV type-specific prevalence using a urine assay in unvaccinated male and female 11- to 18-year olds in Scotland. *Br J Cancer.* 2011;104(7):1221-1226. doi:10.1038/bjc.2011.30
70. Cuschieri K, Kavanagh K, Moore C, Bhatia R, Love J, Pollock KG. Impact of partial bivalent HPV vaccination on vaccine-type infection: A population-based analysis. *Br J Cancer.* 2016;114(11):1261-1264. doi:10.1038/bjc.2016.97
71. Cameron RL, Kavanagh K, Watt C, et al. The impact of bivalent HPV vaccine on cervical intraepithelial neoplasia by deprivation in Scotland: Reducing the gap. *J Epidemiol Community Health (1978).* 2017;71(10):954-960. doi:10.1136/jech-2017-209113
72. Kavanagh K, Pollock KG, Potts A, et al. Introduction and sustained high coverage of the HPV bivalent vaccine leads to a reduction in prevalence of HPV 16/18 and closely related HPV types. *Br J Cancer.* 2014;110(11):2804-2811. doi:10.1038/bjc.2014.198
73. Palmer T, Wallace L, Pollock KG, et al. Prevalence of cervical disease at age 20 after immunisation with bivalent HPV vaccine at age 12-13 in Scotland: Retrospective population study. *BMJ (Online).* 2019;365. doi:10.1136/bmj.l1161
74. Kavanagh K, Sinka K, Cuschieri K, et al. Estimation of HPV prevalence in young women in Scotland; monitoring of future vaccine impact. *BMC Infect Dis.* 2013;13(1). doi:10.1186/1471-2334-13-519
75. Cameron RL, Kavanagh K, Pan J, et al. Human papillomavirus prevalence and herd immunity after introduction of vaccination program, Scotland, 2009–2013. *Emerg Infect Dis.* 2016;22(1):56-64. doi:10.3201/eid2201.150736
76. Kavanagh K, Pollock KG, Cuschieri K, et al. Changes in the prevalence of human papillomavirus following a national bivalent human papillomavirus vaccination programme in Scotland: a 7-year cross-sectional study. *Lancet Infect Dis.* 2017;17(12):1293-1302. doi:10.1016/S1473-3099(17)30468-1

77. Cruickshank M, Pan J, Cotton SC, et al. Reduction in colposcopy workload and associated clinical activity following human papillomavirus (HPV) catch-up vaccination programme in Scotland: an ecological study. *BJOG*. 2017;124(9):1386-1393. doi:10.1111/1471-0528.14562
78. Palmer TJ, McFadden M, Pollock KG, et al. HPV immunisation and cervical screening-confirmation of changed performance of cytology as a screening test in immunised women: A retrospective population-based cohort study. *Br J Cancer*. 2016;114(5):582-589. doi:10.1038/bjc.2015.474
79. Pollock KG, Kavanagh K, Potts A, et al. Reduction of low- and high-grade cervical abnormalities associated with high uptake of the HPV bivalent vaccine in Scotland. *Br J Cancer*. 2014;111(9):1824-1830. doi:10.1038/bjc.2014.479
80. Conway DI, Robertson C, Gray H, et al. Human Papilloma Virus (HPV) Oral Prevalence in Scotland (HOPSCOTCH): A Feasibility Study in Dental Settings. *PLoS One*. 2016;11(11):1-18. doi:10.1371/journal.pone.0165847
81. Schache AG, Powell NG, Cuschieri KS, et al. HPV-Related Oropharynx Cancer in the United Kingdom: An Evolution in the Understanding of Disease Etiology. *Cancer Res*. 2016;76(22):6598. doi:10.1158/0008-5472.CAN-16-0633
82. NHS Inform. Cervical screening (smear test) in Scotland | NHS inform. Published 2020. Accessed February 15, 2023. <https://www.nhsinform.scot/healthy-living/screening/cervical/cervical-screening-smear-test>
83. Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. *Int J Epidemiol*. 2017;46(3):924-938. doi:10.1093/ije/dyw276
84. ISD Scotland. About Information Services Division | ISD Scotland. Accessed September 9, 2021. <https://www.isdscotland.org/About-ISD/>
85. Steinau M, Patel SS, Unger ER. Efficient DNA extraction for HPV genotyping in formalin-fixed, paraffin-embedded tissues. *J Mol Diagn*. 2011;13(4):377-381. doi:10.1016/J.JMOLDX.2011.03.007
86. International Agency for Research on Cancer. IARC Publications Website - Biological Agents. World Health Organization (WHO). Published 2012. Accessed September 7, 2021. <https://publications.iarc.fr/Book-And-Report-Series/Iarc-Monographs-On-The-Identification-Of-Carcinogenic-Hazards-To-Humans/Biological-Agents-2012>
87. Denton K. The proposed BSCC terminology for abnormal cervical cytology. *Cytopathology*. 2008;19(6):398-399. doi:10.1111/j.1365-2303.2008.00624.x
88. Public Health England. Cervical screening: programme and colposcopy management. Guidelines for commissioners, screening providers and programme managers for NHS cervical screening. Gov.uk. Published 2020. Accessed September 7, 2021. <https://www.gov.uk/government/publications/cervical-screening-programme-and-colposcopy-management>
89. Marth C, Landoni F, Mahner S, McCormack M, Gonzalez-Martin A, Colombo N. Cervical cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*. 2017;28(suppl\_4):iv72-iv83. doi:10.1093/annonc/mdx220
90. National Records of Scotland. Population Estimates Time Series Data. Population Estimates Time Series Data. Published 2016. Accessed January 10, 2022. <http://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by->

theme/population/population-estimates/mid-year-population-estimates/population-estimates-time-series-data

91. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
92. Anders S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Soil. Published 2010. Accessed April 18, 2022. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
93. PaVE. Accessed October 31, 2022. [https://pave.niaid.nih.gov/explore/reference\\_genomes/human\\_genomes](https://pave.niaid.nih.gov/explore/reference_genomes/human_genomes)
94. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324
95. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
96. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509
97. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-3066. doi:10.1093/nar/gkf436
98. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033
99. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. 2012;28(4):464-469. doi:10.1093/bioinformatics/btr703
100. García-Alcalde F, Okonechnikov K, Carbonell J, et al. Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*. 2012;28(20):2678-2679. doi:10.1093/bioinformatics/bts503
101. Borgan Ø. Modeling Survival Data: Extending the Cox Model. Terry M. Therneau and Patricia M. Grambsch, Springer-Verlag, New York, 2000. No. of pages: xiii + 350. Price: \$69.95. ISBN 0-387-98784-3. *Stat Med*. 2001;20(13):2053-2054. doi:10.1002/SIM.956
102. Barnier J, Briatte F, Larmarange J. questionr: Functions to make surveys processing easier. *Cran*. Published online 2020. <https://juba.github.io/questionr/>
103. Cuschieri K, Brewster DH, Williams ARW, et al. Distribution of HPV types associated with cervical cancers in Scotland and implications for the impact of HPV vaccines. *Br J Cancer*. 2010;102(5):930-932. doi:10.1038/SJ.BJC.6605556
104. *Globocan. Oropharynx.*; 2020. Accessed September 11, 2021. <https://gco.iarc.fr/today>
105. Schache AG, Powell NG, Cuschieri K, et al. HPV-Related Oropharynx Cancer in the United Kingdom: An Evolution in the Understanding of Disease Etiology. *Cancer Res*. 2016;76(22):6598-6606. doi:10.1158/0008-5472.CAN-16-0633
106. Wells LAR, Junor EJ, Conn B, Pattle S, Cuschieri K. Population-based p16 and HPV positivity rates in oropharyngeal cancer in Southeast Scotland. *J Clin Pathol*. 2015;68(10):849-852. doi:10.1136/jclinpath-2015-202947



107. Wakeham K, Pan J, Pollock KG, Millan D, Bell S, et al. A Prospective Cohort Study of Human Papillomavirus-Driven Oropharyngeal Cancers: Implications for Prognosis and Immunisation. *Clin Oncol (R Coll Radiol)*. 2019;31(9):e132-e142. doi:10.1016/J.CLON.2019.05.010
108. Mesher D, Cuschieri K, Hibbitts S, et al. Type-specific HPV prevalence in invasive cervical cancer in the UK prior to national HPV immunisation programme: Baseline for monitoring the effects of immunisation. *J Clin Pathol*. 2015;68(2):135-140. doi:10.1136/jclinpath-2014-202681
109. Schache AG, Powell NG, Cuschieri K, et al. HPV-Related Oropharynx Cancer in the United Kingdom: An Evolution in the Understanding of Disease Etiology. *Cancer Res*. 2016;76(22):6598-6606. doi:10.1158/0008-5472.CAN-16-0633
110. Hammer A, Rositch A, Qeadan F, Gravitt PE, Blaakaer J. Age-specific prevalence of HPV16/18 genotypes in cervical cancer: A systematic review and meta-analysis. *Int J Cancer*. 2016;138(12):2795-2803. doi:10.1002/ijc.29959
111. De Sanjosé S, Wheeler CM, Quint W, et al. Age-specific occurrence of HPV16- and HPV18-related cervical cancer. *Cancer Epidemiology Biomarkers and Prevention*. 2013;22(7):1313-1318. doi:10.1158/1055-9965.EPI-13-0053
112. Xing B, Guo J, Sheng Y, Wu G, Zhao Y. Human Papillomavirus-Negative Cervical Cancer: A Comprehensive Review. *Front Oncol*. 2021;10. doi:10.3389/FONC.2020.606335
113. Jenkins D, Molijn A, Kazem S, et al. Molecular and pathological basis of HPV-negative cervical adenocarcinoma seen in a global study. *Int J Cancer*. 2020;147(9):2526-2536. doi:10.1002/ijc.33124
114. Petti S. Lifestyle risk factors for oral cancer. *Oral Oncol*. 2009;45(4-5):340-350. doi:10.1016/j.oraloncology.2008.05.018
115. Bradley A. Schiff. Oral Squamous Cell Carcinoma - Ear, Nose, and Throat Disorders - MSD Manual Professional Edition. MSDManual. Published 2022. Accessed February 3, 2022. <https://www.msmanuals.com/en-gb/professional/ear,-nose,-and-throat-disorders/tumors-of-the-head-and-neck/oral-squamous-cell-carcinoma>
116. Arbyn M, Xu L, Simoens C, Martin-Hirsch PPL. Prophylactic vaccination against human papillomaviruses to prevent cervical cancer and its precursors. *Cochrane Database of Systematic Reviews*. 2018;2018(5). doi:10.1002/14651858.CD009069.pub3
117. Liederbach E, Kyrillos A, Wang CH, Liu JC, Sturgis EM, Bhayani MK. The national landscape of human papillomavirus-associated oropharynx squamous cell carcinoma. *Int J Cancer*. 2017;140(3):504-512. doi:10.1002/ijc.30442
118. Lechner M, Liu J, Masterson L, Fenton TR. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nat Rev Clin Oncol*. 2022;19(5):306-327. doi:10.1038/s41571-022-00603-7
119. Lin C, Franceschi S, Clifford GM. Human papillomavirus types from infection to cancer in the anus, according to sex and HIV status: a systematic review and meta-analysis. *Lancet Infect Dis*. 2018;18(2):198-206. doi:10.1016/S1473-3099(17)30653-9
120. Hartwig S, St Guily JL, Dominiak-Felden G, Alemany L, De Sanjosé S. Estimation of the overall burden of cancers, precancerous lesions, and genital warts attributable to 9-valent HPV vaccine types in women and men in Europe. *Infect Agent Cancer*. 2017;12(1):19. doi:10.1186/s13027-017-0129-6

121. Valvo F, Ciurlia E, Avuzzi B, et al. Cancer of the anal region. *Crit Rev Oncol Hematol*. 2019;135:115-127. doi:10.1016/j.critrevonc.2018.12.007
122. Lin C, Slama J, Gonzalez P et al. Cervical determinants of anal HPV infection and high-grade anal lesions in women: a collaborative pooled analysis. *Lancet Infect Dis*. 2019;19(8):880-891. doi:10.1016/S1473-3099(19)30164-1
123. Tseng HF, Morgenstern H, Mack TM, Peters RK. Risk factors for anal cancer: Results of a population-based case-control study. *Cancer Causes and Control*. 2003;14(9):837-846. doi:10.1023/B:CACO.0000003837.10664.7f
124. Dandapani S V, Eaton M, Thomas CR, Pagnini PG. HIV- positive anal cancer: an update for the clinician. *J Gastrointest Oncol*. 2010;1(1):34-44. doi:10.3978/j.issn.2078-6891.2010.005
125. Clifford GM, Georges D, Shiels MS, et al. A meta-analysis of anal cancer incidence by risk group: Toward a unified anal cancer risk scale. *Int J Cancer*. 2021;148(1):38-47. doi:10.1002/ijc.33185
126. Saleem AM, Paulus JK, Shapter AP, Baxter NN, Roberts PL, Ricciardi R. Risk of anal cancer in a cohort with human papillomavirus-related gynecologic neoplasm. *Obstetrics and gynecology*. 2011;117(3):643-649. doi:10.1097/AOG.0B013E31820BFB16
127. Lin C, Slama J, Gonzalez P, et al. Cervical determinants of anal HPV infection and high-grade anal lesions in women: a collaborative pooled analysis. *Lancet Infect Dis*. 2019;19(8):880-891. doi:10.1016/S1473-3099(19)30164-1
128. Saleem AM, Paulus JK, Shapter AP, Baxter NN, Roberts PL, Ricciardi R. Risk of anal cancer in a cohort with human papillomavirus-related gynecologic neoplasm. *Obstetrics and Gynecology*. 2011;117(3):643-649. doi:10.1097/AOG.0b013e31820bfb16
129. Schofield AM, Sadler L, Nelson L, et al. A prospective study of anal cancer screening in HIV-positive and negative MSM. *Aids*. 2016;30(9):1375-1383. doi:10.1097/QAD.0000000000001045
130. Palefsky JM, Holly EA, Ralston ML, Da Costa M, Greenblatt RM. Prevalence and risk factors for anal human papillomavirus infection in human immunodeficiency virus (HIV)-positive and high-risk HIV-negative women. *Journal of Infectious Diseases*. 2001;183(3):383-391. doi:10.1086/318071
131. De Sanjosé S, Serrano B, Tous S, et al. Burden of Human Papillomavirus (HPV)-Related Cancers Attributable to HPVs 6/11/16/18/31/33/45/52 and 58. *JNCI Cancer Spectr*. 2018;2(4):pky045-pky045. doi:10.1093/JNCICS/PKY045
132. International Agency for Research on Cancer. Prostate Source: Globocan 2020 Number of new cases in 2020, both sexes, all ages. Published online 2020:1-2. Accessed September 11, 2021. <https://gco.iarc.fr/today>
133. GLobocan. Cervix uteri. *IARC Globocan*. Published online 2020. Accessed September 11, 2021. <https://gco.iarc.fr/today>
134. Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. *Int J Epidemiol*. 2017;46(3):924-938. doi:10.1093/ije/dyw276
135. Robinson D, Coupland V, Møller H. An analysis of temporal and generational trends in the incidence of anal and other HPV-related cancers in Southeast England. *Br J Cancer*. 2009;100(3):527-531. doi:10.1038/sj.bjc.6604871
136. Brewster DH, Bhatti LA. Increasing incidence of squamous cell carcinoma of the anus in Scotland, 1975–2002. *Br J Cancer*. 2006;95(1):87. doi:10.1038/SJ.BJC.6603175

137. Wakeham K, Kavanagh K. The Burden of HPV-Associated Anogenital Cancers. *Curr Oncol Rep*. 2014;16(9):1-11. doi:10.1007/s11912-014-0402-4
138. Grulich AE, Poynten IM, MacHalek DA, Jin F, Templeton DJ, Hillman RJ. The epidemiology of anal cancer. *Sex Health*. 2012;9(6):504-508. doi:10.1071/SH12070
139. Urbute A, Rasmussen CL, Belmonte F, et al. Prognostic Significance of HPV DNA and p16INK4a in Anal Cancer: A Systematic Review and Meta-Analysis. *Cancer Epidemiology Biomarkers & Prevention*. 2020;29(4):703-710. doi:10.1158/1055-9965.epi-19-1259
140. Baricevic I, He X, Chakrabarty B, et al. High-sensitivity human papilloma virus genotyping reveals near universal positivity in anal squamous cell carcinoma: Different implications for vaccine prevention and prognosis. *Eur J Cancer*. 2015;51(6):776-785. doi:10.1016/j.ejca.2015.01.058
141. Gilbert DC, Williams A, Allan K, et al. P16INK4A, p53, EGFR expression and KRAS mutation status in squamous cell cancers of the anus: Correlation with outcomes following chemo-radiotherapy. *Radiotherapy and Oncology*. 2013;109(1):146-151. doi:10.1016/j.radonc.2013.08.002
142. Wakeham K, Kavanagh K, Cuschieri K, et al. HPV status and favourable outcome in vulvar squamous cancer. *Int J Cancer*. 2017;140(5):1134-1146. doi:10.1002/ijc.30523
143. Sand FL, Rasmussen CL, Frederiksen MH, Andersen KK, Kjaer SK. Prognostic Significance of HPV and p16 Status in Men Diagnosed with Penile Cancer: A Systematic Review and Meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*. 2018;27(10):1123-1132. doi:10.1158/1055-9965.epi-18-0322
144. Rahimi S. HPV-related squamous cell carcinoma of oropharynx: A review. *J Clin Pathol*. 2020;73(10):624-629. doi:10.1136/jclinpath-2020-206686
145. Rodríguez-Carunchio L, Soveral I, Steenbergen RDM, et al. HPV-negative carcinoma of the uterine cervix: A distinct type of cervical cancer with poor prognosis. *BJOG*. 2015;122(1):119-127. doi:10.1111/1471-0528.13071
146. Nicolas I, Marimon L, Barnadas E, et al. HPV-negative tumors of the uterine cervix. *Modern Pathology*. 2019;32(8):1189-1196. doi:10.1038/s41379-019-0249-1
147. King EM, Gilson R, Beddows S, et al. Human papillomavirus DNA in men who have sex with men: type-specific prevalence, risk factors and implications for vaccination strategies. *Br J Cancer*. 2015;112(9):1585-1593. doi:10.1038/BJC.2015.90
148. Kim JY, Park S, Nam BH, et al. Low initial human papilloma viral load implicates worse prognosis in patients with uterine cervical cancer treated with radiotherapy. *J Clin Oncol*. 2009;27(30):5088-5093. doi:10.1200/JCO.2009.22.4659
149. Deng T, Feng Y, Zheng J, Huang Q, Liu J. Low initial human papillomavirus viral load may indicate worse prognosis in patients with cervical carcinoma treated with surgery. *J Gynecol Oncol*. 2015;26(2):111-117. doi:10.3802/JGO.2015.26.2.111
150. Stevenson A, Wakeham K, Pan J, et al. Droplet digital PCR quantification suggests that higher viral load correlates with improved survival in HPV-positive oropharyngeal tumours. *J Clin Virol*. 2020;129. doi:10.1016/J.JCV.2020.104505
151. Cohen MA, Basha SR, Reichenbach DK, Robertson E, Sewell DA. Increased viral load correlates with improved survival in HPV-16-associated tonsil carcinoma patients. *Acta Otolaryngol*. 2008;128(5):583-589. doi:10.1080/00016480701558880

152. Mellin H, Dahlgren L, Munck-Wikland E, et al. Human papillomavirus type 16 is episomal and a high viral load may be correlated to better prognosis in tonsillar cancer. *Int J Cancer*. 2002;102(2):152-158. doi:10.1002/IJC.10669
153. Hashida Y, Higuchi T, Matsumoto S, et al. Prognostic significance of human papillomavirus 16 viral load level in patients with oropharyngeal cancer. *Cancer Sci*. 2021;112(10):4404-4417. doi:10.1111/cas.15105
154. Deng T, Feng Y, Zheng J, Huang Q, Liu J. Low initial human papillomavirus viral load may indicate worse prognosis in patients with cervical carcinoma treated with surgery. *J Gynecol Oncol*. 2015;26(2):111-117. doi:10.3802/JGO.2015.26.2.111
155. Biesaga B, Mucha-Matecka A, Janecka-Widła A, et al. Differences in the prognosis of HPV16-positive patients with squamous cell carcinoma of head and neck according to viral load and expression of P16. *J Cancer Res Clin Oncol*. 2018;144(1):63-73. doi:10.1007/S00432-017-2531-2
156. Małusecka E, Chmielik E, Suwiński R, et al. Significance of HPV16 Viral Load Testing in Anal Cancer. *Pathol Oncol Res*. 2020;26(4):2191-2199. doi:10.1007/S12253-020-00801-7
157. Poizot-Martin I, Henry M, Benhaim S, Obry-Roguet V, Figarella D, Tamalet C. High level of HPV 16 and 18 DNA load in anal swabs from male and female HIV-1 infected patients. *J Clin Virol*. 2009;44(4):314-317. doi:10.1016/J.JCV.2009.02.003
158. Pierangeli A, Scagnolari C, Degener AM, et al. Type-specific human papillomavirus-DNA load in anal infection in HIV-positive men. *AIDS*. 2008;22(15):1929-1935. doi:10.1097/QAD.0B013E32830FBD7A
159. Drobacheff C, Dupont P, Mouglin C, et al. Anal human papillomavirus DNA screening by Hybrid Capture II™ in human immunodeficiency virus-positive patients with or without anal intercourse. *European Journal of Dermatology*. 2003;13(4):367-371. Accessed September 11, 2021. [http://www.jle.com/en/revues/ejd/e-docs/anal\\_human\\_papillomavirus\\_dna\\_screening\\_by\\_hybrid\\_capture\\_ii\\_tm\\_in\\_human\\_immunodeficiency\\_virus\\_positive\\_pati\\_260502/article.phtml?tab=texte](http://www.jle.com/en/revues/ejd/e-docs/anal_human_papillomavirus_dna_screening_by_hybrid_capture_ii_tm_in_human_immunodeficiency_virus_positive_pati_260502/article.phtml?tab=texte)
160. Rödel F, Wieland U, Fraunholz I, et al. Human papillomavirus DNA load and p16INK4a expression predict for local control in patients with anal squamous cell carcinoma treated with chemoradiotherapy. *Int J Cancer*. 2015;136(2):278-288. doi:10.1002/IJC.28979
161. Urbute A, Rasmussen CL, Belmonte F, et al. Prognostic Significance of HPV DNA and p16INK4a in Anal Cancer: A Systematic Review and Meta-Analysis. *Cancer Epidemiology and Prevention Biomarkers*. 2020;29(4):703-710. doi:10.1158/1055-9965.EPI-19-1259
162. Finzer P, Aguilar-Lemarroy A, Rösl F. The role of human papillomavirus oncoproteins E6 and E7 in apoptosis. *Cancer Lett*. 2002;188(1-2):15-24. doi:10.1016/S0304-3835(02)00431-7
163. Jeon S, Lambert PF. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc Natl Acad Sci U S A*. 1995;92(5):1654. doi:10.1073/PNAS.92.5.1654
164. Glynne-Jones R, Nilsson PJ, Aschele C, et al. Anal cancer: ESMO-ESSO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*. 2014;25(3):10-20. doi:10.1093/annonc/mdu159
165. SS N, SL M, SK J. Quality of Life After Radiotherapy for Rectal and Anal Cancer. *Curr Colorectal Cancer Rep*. 2020;16(1). doi:10.1007/S11888-019-00448-W

166. Fakhrian K, Sauer T, Dinkel A, et al. Chronic adverse events and quality of life after radiochemotherapy in anal cancer patients: A single institution experience and review of the literature. *Strahlentherapie und Onkologie*. 2013;189(6):486-494. doi:10.1007/s00066-013-0314-5
167. Das P, Cantor SB, Parker CL, et al. Long-term quality of life after radiotherapy for the treatment of anal cancer. *Cancer*. 2010;116(4):822-829. doi:10.1002/cncr.24906
168. Neibart SS, Manne SL, Jabbour SK. Quality of Life After Radiotherapy for Rectal and Anal Cancer. *Curr Colorectal Cancer Rep*. 2020;16(1). doi:10.1007/S11888-019-00448-W
169. Stevenson A, Wakeham K, Pan J, et al. Droplet digital PCR quantification suggests that higher viral load correlates with improved survival in HPV-positive oropharyngeal tumours. *Journal of Clinical Virology*. 2020;129:104505. doi:10.1016/j.jcv.2020.104505
170. Guerendiain D, Grigorescu R, Kirk A, et al. HPV status and HPV16 viral load in anal cancer and its association with clinical outcome. *Cancer Med*. 2022;11(22):4193-4203. doi:10.1002/cam4.4771
171. Anon. Anal cancer incidence statistics | Cancer Research UK. Cancer Research UK. Published 2015. Accessed September 11, 2021. <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/anal-cancer/incidence#heading-Zero>
172. Moscicki AB, Darragh TM, Michael Berry-Lawhorn J, et al. Screening for anal cancer in women. *J Low Genit Tract Dis*. 2015;19(3):S27-S42. doi:10.1097/LGT.0000000000000117
173. Li P, Tan Y, Zhu LX, et al. Prognostic value of HPV DNA status in cervical cancer before treatment: A systematic review and meta-analysis. *Oncotarget*. 2017;8(39):66352-66359. doi:10.18632/oncotarget.18558
174. Nicolás I, Marimon L, Barnadas E, et al. HPV-negative tumors of the uterine cervix. *Mod Pathol*. 2019;32(8):1189-1196. doi:10.1038/S41379-019-0249-1
175. Wakeham K, Pan J, Pollock KG, et al. A Prospective Cohort Study of Human Papillomavirus-Driven Oropharyngeal Cancers: Implications for Prognosis and Immunisation. *Clin Oncol (R Coll Radiol)*. 2019;31(9):e132-e142. doi:10.1016/J.CLON.2019.05.010
176. Ang KK, Harris J, Wheeler R, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *New England Journal of Medicine*. 2010;363(1):24-35. doi:10.1056/nejmoa0912217
177. Sand FL, Rasmussen CL, Frederiksen MH, Andersen KK, Kjaer SK. Prognostic Significance of HPV and p16 Status in Men Diagnosed with Penile Cancer: A Systematic Review and Meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2018;27(10):1123-1132. doi:10.1158/1055-9965.EPI-18-0322
178. Chu C, Chen K, Tan X, et al. Prevalence of human papillomavirus and implication on survival in Chinese penile cancer. *Virchows Archiv*. 2020;477(5):667-675. doi:10.1007/s00428-020-02831-7
179. Rasmussen CL, Sand FL, Hoffmann Frederiksen M, Kaae Andersen K, Kjær SK. Does HPV status influence survival after vulvar cancer? *Int J Cancer*. 2018;142(6):1158-1165. doi:10.1002/ijc.31139
180. Mellin H, Dahlgren L, Munck-Wikland E, et al. Human papillomavirus type 16 is episomal and a high viral load may be correlated to better prognosis in tonsillar cancer. *Int J Cancer*. 2002;102(2):152-158. doi:10.1002/IJC.10669

181. Veitía D, Liuzzi J, Ávila M, Rodriguez I, Toro F, Correnti M. Association of viral load and physical status of HPV-16 with survival of patients with head and neck cancer. *Ecancermedicalscience*. 2020;14. doi:10.3332/ECANCER.2020.1082
182. Olthof NC, Straetmans JMJA, Snoeck R, Ramaekers FCS, Kremer B, Speel EJM. Next-generation treatment strategies for human papillomavirus-related head and neck squamous cell carcinoma: Where do we go? *Rev Med Virol*. 2012;22(2):88-105. doi:10.1002/rmv.714
183. Olthof NC, Huebbers CU, Kolligs J, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. *Int J Cancer*. 2015;136(5):E207-E218. doi:10.1002/ijc.29112
184. Theophanous S, Samuel R, Lilley J, et al. Prognostic factors for patients with anal cancer treated with conformal radiotherapy-a systematic review. *BMC Cancer*. 2022;22(1). doi:10.1186/S12885-022-09729-4
185. Schiffman M, Doorbar J, Wentzensen N, et al. Carcinogenic human papillomavirus infection. *Nat Rev Dis Primers*. 2016;2. doi:10.1038/nrdp.2016.86
186. Haile S, Corbett RD, Bilobram S, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res*. 2019;47(2):e12. doi:10.1093/nar/gky1142
187. McDonough SJ, Bhagwate A, Sun Z, et al. Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. *PLoS One*. 2019;14(4):1-15. doi:10.1371/journal.pone.0211400
188. Božić L, Jovanović T, Šmitran A, Janković M, Knežević A. Comparison of HPV detection rate in formalin-fixed paraffin-embedded tissues of head and neck carcinoma using two DNA extraction kits and three amplification methods. *Eur J Oral Sci*. 2020;128(6):501-507. doi:10.1111/eos.12746
189. Harlé A, Guillet J, Thomas J, et al. Evaluation and validation of HPV real-time PCR assay for the detection of HPV DNA in oral cytobrush and FFPE samples. *Sci Rep*. 2018;8(1). doi:10.1038/s41598-018-29790-z
190. Dal Bello B, Spinillo A, Alberizzi P, Cesari S, Gardella B, Silini EM. Validation of the SPF10 LiPA human papillomavirus typing assay using formalin-fixed paraffin-embedded cervical biopsy samples. *J Clin Microbiol*. 2009;47(7):2175-2180. doi:10.1128/JCM.00286-09
191. Guerendiain D, Moore C, Wells LAR, Conn B, Cuschieri K. Formalin fixed paraffin embedded (FFPE) material is amenable to HPV detection by the Xpert® HPV assay. *Journal of Clinical Virology*. 2016;77:55-59. doi:10.1016/j.jcv.2016.02.007
192. Lagheden C, Eklund C, Kleppe SN, Unger ER, Dillner J, Sundström K. Validation of a standardized extraction method for formalin-fixed paraffin-embedded tissue samples. *Journal of Clinical Virology*. 2016;80:36-39. doi:10.1016/j.jcv.2016.04.016
193. Kerr DA, Sweeney B, Arpin RN, et al. Automated extraction of formalin-fixed, paraffin-embedded tissue for high-risk human papillomavirus testing of head and neck squamous cell carcinomas using the roche cobas 4800 system. *Arch Pathol Lab Med*. 2016;140(8):844-848. doi:10.5858/arpa.2015-0272-OA
194. Roberts CC, Swoyer R, Bryan JT, Taddeo FJ. Comparison of real-time multiplex Human Papillomavirus (HPV) PCR assays with the linear array HPV genotyping PCR assay and influence of DNA extraction method on HPV detection. *J Clin Microbiol*. 2011;49(5):1899-1906. doi:10.1128/JCM.00235-10

195. Atchison S, Shilling H, Balgovind P, et al. Evaluation of the Roche MagNA Pure 96 nucleic acid extraction platform for the Seegene Anyplex II HPV28 detection assay. *J Appl Microbiol.* 2021;131(5):2592-2599. doi:10.1111/jam.15126
196. Long-read sequencing vs short-read sequencing. Accessed May 1, 2023. <https://frontlinegenomics.com/long-read-sequencing-vs-short-read-sequencing/>
197. Ottestad AL, Emdal EF, Grønberg BH, Halvorsen TO, Dai HY. Fragmentation assessment of FFPE DNA helps in evaluating NGS library complexity and interpretation of NGS results. *Exp Mol Pathol.* 2022;126:104771. doi:10.1016/J.YEXMP.2022.104771
198. Hao Y, Yang L, Galvao Neto A, et al. HPVViewer: Sensitive and specific genotyping of human papillomavirus in metagenomic DNA. *Bioinformatics.* 2018;34(12):1986-1995. doi:10.1093/bioinformatics/bty037
199. Chandrani P, Kulkarni V, Iyer P, et al. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br J Cancer.* 2015;112(12):1958-1965. doi:10.1038/bjc.2015.121
200. Arroyo Mühr LS, Lagheden C, Hultin E, et al. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study. *Br J Cancer.* 2018;119(9):1163-1168. doi:10.1038/s41416-018-0311-7
201. Hildesheim A, Schiffman M, Bromley C, et al. Human papillomavirus type 16 variants and risk of cervical cancer. *J Natl Cancer Inst.* 2001;93(4):315-318. doi:10.1093/JNCI/93.4.315
202. M SS, AC R, Z C, et al. A population-based prospective study of carcinogenic human papillomavirus variant lineages, viral persistence, and cervical neoplasia. *Cancer Res.* 2010;70(8):3159-3169. doi:10.1158/0008-5472.CAN-09-4179
203. Nicolás-Párraga S, Gandini C, Pimenoff VN, et al. HPV16 variants distribution in invasive cancers of the cervix, vulva, vagina, penis, and anus. *Cancer Med.* 2016;5(10):2909-2919. doi:10.1002/cam4.870
204. Volpini LPB, Boldrini NAT, De Freitas LB, Miranda AE, Spano LC. The high prevalence of HPV and HPV16 European variants in cervical and anal samples of HIV-seropositive women with normal Pap test results. *PLoS One.* 2017;12(4). doi:10.1371/journal.pone.0176422
205. Ferreira MT, Giulia Gonçalves M, Mendoza López RV, Sichero L. Genetic variants of HPV-16 and their geographical and anatomical distribution in men: A systematic review with meta-analysis. *Virology.* 2021;558:134-144. doi:10.1016/j.virol.2021.03.007
206. van der Weele P, Meijer CJLM, King AJ. Whole-Genome Sequencing and Variant Analysis of Human Papillomavirus 16 Infections. *J Virol.* 2017;91(19). doi:10.1128/jvi.00844-17
207. Van Der Weele P, Meijer CJLM, King AJ. High whole-genome sequence diversity of human papillomavirus type 18 isolates. *Viruses.* 2018;10(2):6-19. doi:10.3390/v10020068
208. Zehbe I, Tachezy R, Mytilineos J, et al. Human papillomavirus 16 E6 polymorphisms in cervical lesions from different European populations and their correlation with human leukocyte antigen class II haplotypes. *Int J Cancer.* 2001;94(5):711-716. doi:10.1002/IJC.1520
209. Gheit T, Cornet I, Clifford GM, et al. Risks for persistence and progression by human papillomavirus type 16 variant lineages among a population-based sample of Danish women. *Cancer Epidemiol Biomarkers Prev.* 2011;20(7):1315-1321. doi:10.1158/1055-9965.EPI-10-1187

210. Cañadas MP, Darwich L, Sirera G, et al. Human papillomavirus 16 integration and risk factors associated in anal samples of HIV-1 infected men. *Sex Transm Dis*. 2010;37(5):311-315. doi:10.1097/OLQ.0b013e3181c9c23f
211. Morel A, Neuzillet C, Wack M, et al. Mechanistic signatures of human papillomavirus insertions in anal squamous cell carcinomas. *Cancers (Basel)*. 2019;11(12). doi:10.3390/cancers11121846
212. Jeannot E, Harlé A, Holmes A, Sastre-Garau X. Nuclear factor I X is a recurrent target for HPV16 insertions in anal carcinomas. *Genes Chromosomes Cancer*. 2018;57(12):638-644. doi:10.1002/gcc.22675
213. Lagström S, Løvestad AH, Umu SU, et al. HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples. *Tumour Virus Res*. 2021;12. doi:10.1016/j.tvr.2021.200221
214. Allaire J. RStudio: integrated development for R. *RStudio Team*. Published online 2012: <http://www.rstudio.com/>. Accessed April 11, 2022. <http://www.rstudio.com/>
215. Gonçalves MG, Ferreira MT, López RVM, et al. Prevalence and persistence of HPV-16 molecular variants in the anal canal of men: The HIM study. *Journal of Clinical Virology*. 2022;149. doi:10.1016/j.jcv.2022.105128
216. Lang Kuhs KA, Faden DL, Chen L, et al. Genetic variation within the human papillomavirus type 16 genome is associated with oropharyngeal cancer prognosis. *Annals of Oncology*. 2022;33(6):638-648. doi:10.1016/j.annonc.2022.03.005
217. Godi A, Kemp TJ, Pinto LA, Beddows S. Sensitivity of Human Papillomavirus (HPV) Lineage and Sublineage Variant Pseudoviruses to Neutralization by Nonavalent Vaccine Antibodies. *Journal of Infectious Diseases*. 2019;220(12):1940-1945. doi:10.1093/infdis/jiz401
218. Brim H, Mirabello L, Bass S, Ford DH, Carethers JM, Ashktorab H. Association of Human Papillomavirus Genotype 16 Lineages With Anal Cancer Histologies Among African Americans. *Gastroenterology*. 2021;160(3):922-924. doi:10.1053/j.gastro.2020.10.022
219. Saraiya M, Steben M, Watson M, Markowitz L. Evolution of cervical cancer screening and prevention in united states and canada: Implications for public health practitioners and clinicians. *Prev Med (Baltim)*. 2013;57(5):426-433. doi:10.1016/j.ypmed.2013.01.020
220. Mühr LSA, Guerendiain D, Cuschieri K, Sundström K. Human papillomavirus detection by whole-genome next-generation sequencing: Importance of validation and quality assurance procedures. *Viruses*. 2021;13(7). doi:10.3390/v13071323
221. Gradissimo A, Burk RD. Molecular tests potentially improving HPV screening and genotyping for cervical cancer prevention. *Expert Rev Mol Diagn*. 2017;17(4):379-391. doi:10.1080/14737159.2017.1293525
222. Arroyo Mühr LS, Bzhalava D, Lagheden C, et al. Does human papillomavirus-negative condylomata exist? *Virology*. 2015;485:283-288. doi:10.1016/j.virol.2015.07.023
223. Arroyo Mühr LS, Hultin E, Bzhalava D, et al. Human papillomavirus type 197 is commonly present in skin tumors. *Int J Cancer*. 2015;136(11):2546-2555. doi:10.1002/ijc.29325
224. Meijer CJLM, Berkhof J, Castle PE, et al. Guidelines for human papillomavirus DNA test requirements for primary cervical cancer screening in women 30 years and older. *Int J Cancer*. 2009;124(3):516-520. doi:10.1002/ijc.24010



225. Brancaccio RN, Robitaille A, Dutta S, Rollison DE, Tommasino M, Gheit T. MinION nanopore sequencing and assembly of a complete human papillomavirus genome. *J Virol Methods*. 2021;294. doi:10.1016/j.jviromet.2021.114180
226. Yang S, Zhao Q, Tang L, et al. Whole Genome Assembly of Human Papillomavirus by Nanopore Long-Read Sequencing. *Front Genet*. 2022;12. doi:10.3389/fgene.2021.798608
227. Yang W, Liu Y, Dong R, et al. Accurate Detection of HPV Integration Sites in Cervical Cancer Samples Using the Nanopore MinION Sequencer Without Error Correction. *Front Genet*. 2020;11. doi:10.3389/fgene.2020.00660
228. Zhou L, Qiu Q, Zhou Q, et al. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. *Nat Commun*. 2022;13(1):1-18. doi:10.1038/s41467-022-30190-1
229. Information Services Division. *Scottish Cervical Screening Programme*. Vol 94.; 2002.
230. Krivacsy S, Bayingana A, Binagwaho A. Affordable human papillomavirus screening needed to eradicate cervical cancer for all. *Lancet Glob Health*. 2019;7(12):e1605-e1606. doi:10.1016/S2214-109X(19)30423-1
231. Godi A, Boamong D, Elegunde B, et al. Comprehensive Assessment of the Antigenic Impact of Human Papillomavirus Lineage Variation on Recognition by Neutralizing Monoclonal Antibodies Raised against Lineage A Major Capsid Proteins of Vaccine-Related Genotypes. *J Virol*. 2020;94(24):1236-1256. doi:10.1128/jvi.01236-20
232. Mattox AK, D'Souza G, Khan Z, et al. Comparison of next generation sequencing, droplet digital PCR, and quantitative real-time PCR for the earlier detection and quantification of HPV in HPV-positive oropharyngeal cancer. *Oral Oncol*. 2022;128. doi:10.1016/j.oraloncology.2022.105805
233. Leung E, Han K, Zou J, et al. HPV sequencing facilitates ultrasensitive detection of HPV circulating tumor DNA. *Clinical Cancer Research*. 2021;27(21):5857-5868. doi:10.1158/1078-0432.CCR-19-2384
234. Falcaro M, Castañon A, Ndlela B, et al. The effects of the national HPV vaccination programme in England, UK, on cervical cancer and grade 3 cervical intraepithelial neoplasia incidence: a register-based observational study. *The Lancet*. 2021;398(10316):2084-2092. doi:10.1016/S0140-6736(21)02178-4
235. Lei J, Ploner A, Elfström KM, et al. HPV Vaccination and the Risk of Invasive Cervical Cancer. *New England Journal of Medicine*. 2020;383(14):1340-1348. doi:10.1056/nejmoa1917338
236. American Joint Committee on Cancer. AJCC Cancer Staging Manual. 8th ed. Anus. *Springer*. Published online 2017:275. Accessed September 11, 2021. www.cancerstaging.orgajcc@facs.org

**10. Appendix 1. HPV status according to demographics and clinical variables in cervical cancer samples. Odds ratio (univariate and adjusted) were calculated for overall HPV positivity.**


able	Level	N	% (I/)	N HPV +	%HPV+ (/N)	N HPV 16+	%HPV 16+ (/N)	N HPV 16/18+	%HPV 16/18+ (/N)	Unadjusted OR (95% Cis) Overall HPV	p value	Adjusted OR (95% Cis) Overall HPV	p value
<b>Age</b>	<b>&lt;45</b>	235	36.66%	226	96.17 (92.88 - 97.97)	157	66.81 (60.56 - 72.52)	191	81.28 (75.80 - 85.75)	1		1	
	<b>45-54</b>	107	16.69%	95	88.79 (81.42 - 93.47)	58	54.21 (44.79 - 63.34)	81	75.70 (66.78 - 82.84)	0.31 (0.12 - 0.77)	0.012	0.34 (0.14 - 0.90)	0.029
	<b>55 - 64</b>	53	8.27%	46	86.79 (75.16 - 93.45)	34	64.15 (50.69 - 75.70)	39	73.58 (60.41 - 83.56)	0.26 (0.09 - 0.76)	0.011	0.32 (0.1 - 1.11)	0.058
	<b>65 - 74</b>	44	6.86%	37	84.09 (70.63 - 92.07)	30	68.18 (53.44 - 80.0)	33	75.0 (60.56 - 85.43)	0.21 (0.07 - 0.62)	0.003	0.19 (0.06 - 0.60)	0.003
	<b>≥75</b>	48	7.49%	39	81.25 (68.06 - 89.81)	24	50.0 (36.39 - 63.61)	28	58.33 (44.28 - 71.15)	0.17 (0.06 - 0.47)	<0.001	0.13 (0.04 - 0.40)	<0.001
	<b>Not available</b>	154	24.02%	144	93.51 (88.46 - 96.44)	91	59.09 (51.20 - 66.54)	118		-	-	-	-
<b>Collection year</b>	<b>2015</b>	244	38.07%	218	89.34 (84.84 - 92.62)	151	61.89 (55.66 - 67.75)	193	79.10 (73.57 - 83.73)	1		1	
	<b>2016</b>	222	34.63%	210	94.59 (90.79 - 96.88)	137	61.71 (55.17 - 67.85)	170	76.58 (70.59 - 81.67)	2.087 (1.04 - 4.388)	0.042	1.86 (0.72 - 5.18)	0.213
	<b>2017</b>	175	27.30%	159	90.86 (85.67 - 94.30)	106	60.57 (53.18 - 67.51)	127	72.57 (65.53 - 78.64)	1.185 (0.621 - 2.325)	0.6113	1.06 (0.41 - 2.84)	0.907
<b>Histology</b>	<b>SCC</b>	356	55.54%	336	94.38 (91.48 - 96.33)	238	66.85 (61.80 - 71.54)	273	76.69 (72.03 - 80.78)	1		1	
	<b>ASC&amp;ADC</b>	122	19.03%	102	83.61 (76.04 - 89.13)	62	50.82 (42.06 - 59.53)	95	77.87 (69.72 - 84.32)	0.30 (0.16 - 0.59)	<0.001	0.30 (0.14 - 0.61)	0.001
	<b>Unknown</b>	163	25.43%	149	91.41 (86.10 - 94.81)	94	57.67 (49.99 - 64.99)	122	74.85 (67.67 - 80.89)	0.074 (0.018 - 0.320)	<0.001	0.08 (0.01 - 0.44)	0.003
<b>Location</b>	<b>NHS Greater Glasgow &amp; Clyde</b>	210	32.76%	196	93.33 (89.12 - 95.99)	127	60.48 (53.74 - 66.85)	161	76.67 (70.50 - 81.88)	1		1	
	<b>NHS Lanarkshire</b>	37	5.77%	37	100 (90.59 - 100)	23	62.16 (46.10 - 75.93)	33	89.19 (75.29 - 95.72)	0.001 (0.000 - NA)	0.982	1.05e+6 (9.26e-7 - 2.25e+137)	0.986
	<b>NHS Tayside</b>	91	14.20%	87	95.60 (89.23 - 98.28)	59	64.84 (54.61 - 73.87)	70	76.92 (67.28 - 84.38)	1.553 (0.539 - 5.601)	0.449	1.14 (0.24 - 5.69)	0.86
	<b>NHS Grampian</b>	34	5.30%	26	76.47 (60.0 - 87.56)	20	58.82 (42.22 - 73.63)	25	73.53 (56.88 - 85.30)	0.232 (0.090 - 0.629)	0.003	0.22 (0.04 - 0.85)	0.032
	<b>NHS Highlands</b>	24	3.74%	18	75.0 (55.10 - 88.0)	15	62.50 (42.71 - 78.84)	18	75.0 (5.51 - 88.0)	0.2143 (0.0754 - 0.665)	0.005	0.12 (0.02 - 0.61)	0.01
	<b>NHS Lothian</b>	112	17.47%	104	92.86 (86.54 - 96.34)	73	65.18 (55.99 - 73.37)	89	79.46 (71.06 - 85.90)	0.928 (0.385 - 2.391)	0.872	0.62 (0.14 - 2.42)	0.501
	<b>NHS Forth Valley</b>	65	10.14%	62	95.38 (87.28 - 98.42)	39	60.0 (47.86 - 71.03)	50	76.92 (65.35 - 85.48)	1.089 (0.374 - 3.952)	0.884	0.65 (0.13 - 3.21)	0.586
<b>NHS Fife</b>	68	10.61%	58	85.29 (75.0 - 91.81)	38	55.88 (44.08 - 67.05)	44	64.71 (52.85 - 75.0)	0.414 (0.176 - 1.007)	0.0453	0.19 (0.04 - 0.73)	0.02	

Appendix 2. HPV status according to demographics in oropharyngeal cancer samples. Odds ratio (univariate and adjusted) were calculated for overall HPV positivity.

Variable	Level	N HPV+ve	% HPV+ve (/N)	N HPV 16+ve	% HPV 16+ve (/N)	Unadjusted OR (95% Cis) Overall HPV	p value	Adjusted OR (95% Cis) Overall HPV	p value	Unadjusted OR (95% Cis) HPV 16+ve	p value	Adjusted OR (95% Cis) HPV 16+ve	p value
Sex	Female (N=479)	215	44.89 (40.49 - 49.37)	192	40.08 (35.79 - 44.53)	1		1		1		1	
	Male (N=1319)	772	58.53 (55.85 - 61.16)	723	54.81 (52.11 - 57.48)	1.73 (1.40 - 0.14)	<0.001	1.68 (1.34 - 2.11)	<0.001	1.70 (1.36 - 2.14)	<0.001	1.63 (1.28 - 2.09)	<0.001
Age group	<50 (N=217)	151	69.59 (63.17 - 75.33)	137	63.13 (56.53 - 69.27)	1		1		1		1	
	50 - 59 (N=590)	395	66.95 (63.06 - 70.63)	376	63.73 (59.77 - 67.51)	0.88 (0.63 - 1.23)	0.478	0.90 (0.63 - 1.27)	0.539	0.94 (0.66 - 1.34)	0.742	0.92 (0.68 - 1.43)	0.965
	60 - 69 (N=593)	299	50.42 (46.41 - 54.43)	274	46.21 (42.23 - 50.23)	0.44 (0.32 - 0.62)	<0.001	0.43 (0.30 - 0.60)	<0.001	0.50 (0.35 - 0.70)	<0.001	0.48 (0.33 - 0.90)	<0.001
	> 70 (N=394)	138	30.03 (30.48 - 39.86)	125	31.73 (27.33 - 36.48)	0.23 (0.16 - 0.33)	<0.001	0.25 (0.17 - 0.36)	<0.001	0.25 (0.17 - 0.37)	<0.001	0.27 (0.18 - 0.40)	<0.001
Location	NHS Greater Glasgow & Clyde (N=868)	419	48.27 (44.96 - 51.59)	377	43.43 (40.17 - 46.75)	1		1		1		1	
	NHS Ayrshire & Arran (N=165)	85	52.52 (43.95 - 59.02)	80	48.48 (40.98 - 56.05)	1.30 (0.92 - 1.84)	0.133	1.37 (0.95 - 1.96)	0.09	1.46 (1.01 - 2.14)	0.045	1.50 (1.02 - 2.22)	0.041
	NHS Dumfries & Galloway (N=4)	40	0 (0 - 8.76)	0	0 (0 - 8.76)	1.37e-6 (NA - 2.95e8)	0.96	1.66e-6 (NA - 5.55e6)	0.958	1.27e-6 (NA - 2.72e7)	0.959	5.99e-7 (NA - 4.11e14)	0.973
	NHS Forth Valley (N=72)	40	55.56 (44.09 - 66.47)	38	2.78 (41.4 - 63.88)	1.33 (0.83 - 2.18)	0.236	1.36 (0.82 - 2.26)	0.233	3.13 (1.66 - 6.35)	<0.001	3.16 (1.64 - 6.53)	<0.001
	NHS Grampian (N=192)	173	90.10 (85.06 - 93.57)	163	84.90 (79.15 - 89.28)	9.76 (6.12 - 16.46)	<0.001	9.61 (5.97 - 16.34)	<0.001	8.06 (5.08 - 13.48)	<0.001	8.21 (5.12 - 13.84)	<0.001
	NHS Highlands (N=122)	75	61.48 (52.62 - 69.64)	71	58.20 (49.33 - 66.57)	1.71 (1.16 - 2.53)	0.007	1.84 (1.13 - 2.77)	0.003	2.13 (1.38 - 3.33)	0.007	2.35 (1.51 - 3.74)	<0.001
	NHS Lanarkshire (N=330)	163	40.39 (44.03 - 54.76)	156	47.27 (41.95 - 52.66)	1.05 (0.82 - 1.36)	0.694	1.05 (0.81 - 1.37)	0.701	1.03 (0.79 - 1.34)	0.833	1.07 (0.81 - 1.41)	0.645
	NHS Lothian (N=56)	32	57.14 (44.13 - 69.23)	30	53.57 (40.7 - 65.98)	1.43 (0.83 - 2.49)	0.2	1.29 (0.74 - 2.29)	0.372	1.19 (0.68 - 2.07)	0.54	1.10 (0.63 - 1.95)	0.736

Article

# Human Papillomavirus Detection by Whole-Genome Next-Generation Sequencing: Importance of Validation and Quality Assurance Procedures

Laila Sara Arroyo Mühr <sup>1,†</sup>, Daniel Guerendiain <sup>2,3,†</sup>, Kate Cuschieri <sup>2,‡</sup> and Karin Sundström <sup>4,\*</sup> 

<sup>1</sup> International HPV Reference Center, Department of Laboratory Medicine, Karolinska Institutet, SE-141 86 Stockholm, Sweden; sara.arroyo.muhr@ki.se

<sup>2</sup> Scottish Human Papillomavirus Reference Laboratory (SHPVRL), Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK; Kate.Cuschieri@nhslothian.scot.nhs.uk

<sup>3</sup> School of Medicine, University of St Andrews, St Andrews KY16 9TF, UK; dgr7@st-andrews.ac.uk

<sup>4</sup> Department of Laboratory Medicine, Karolinska Institutet, SE-141 86 Stockholm, Sweden

\* Correspondence: Karin.sundstrom@ki.se

† Equal contribution.

‡ Equal contribution.



**Citation:** Arroyo Mühr, L.S.; Guerendiain, D.; Cuschieri, K.; Sundström, K. Human Papillomavirus Detection by Whole-Genome Next-Generation Sequencing: Importance of Validation and Quality Assurance Procedures. *Viruses* **2021**, *13*, 1323. <https://doi.org/10.3390/v13071323>

Academic Editors: Lisa Mirabello and Meredith Yeager

Received: 6 May 2021

Accepted: 18 June 2021

Published: 8 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Next-generation sequencing (NGS) yields powerful opportunities for studying human papillomavirus (HPV) genomics for applications in epidemiology, public health, and clinical diagnostics. HPV genotypes, variants, and point mutations can be investigated in clinical materials and described in previously unprecedented detail. However, both the NGS laboratory analysis and bioinformatical approach require numerous steps and checks to ensure robust interpretation of results. Here, we provide a step-by-step review of recommendations for validation and quality assurance procedures of each step in the typical NGS workflow, with a focus on whole-genome sequencing approaches. The use of directed pilots and protocols to ensure optimization of sequencing data yield, followed by curated bioinformatical procedures, is particularly emphasized. Finally, the storage and sharing of data sets are discussed. The development of international standards for quality assurance should be a goal for the HPV NGS community, similar to what has been developed for other areas of sequencing efforts including microbiology and molecular pathology. We thus propose that it is time for NGS to be included in the global efforts on quality assurance and improvement of HPV-based testing and diagnostics.

**Keywords:** human papillomavirus; HPV; next-generation sequencing; NGS; whole-genome sequencing; WGS; deep sequencing

## 1. Introduction

Tests for the detection of human papillomavirus (HPV) infection in humans have evolved dramatically over the last decades. Initial low-throughput hybridization/blotting techniques prefaced broad-spectrum signal amplification assays, which were then replaced by rapid high-throughput target-amplification assays involving quantitative polymerase chain reaction (qPCR). The latter tests are capable of detecting individual HPV-genotypes and have become the mainstay of HPV-based screening and clinical testing [1]. Arguably, the next “age” of HPV testing should involve going beyond simply detecting the presence or absence of HPV but, rather, providing more detailed insight into the likely course and clinical consequences of HPV infection.

Next-generation sequencing (NGS) of human or microbial genetic material is being applied increasingly in laboratory contexts, to facilitate research, population-based epidemiology, and recently, personalized patient diagnostics. NGS can be used as a highly sensitive method for HPV detection due to its ability to detect types at low copy number (even within multiple infections), novel types, and/or known types that are distantly related to

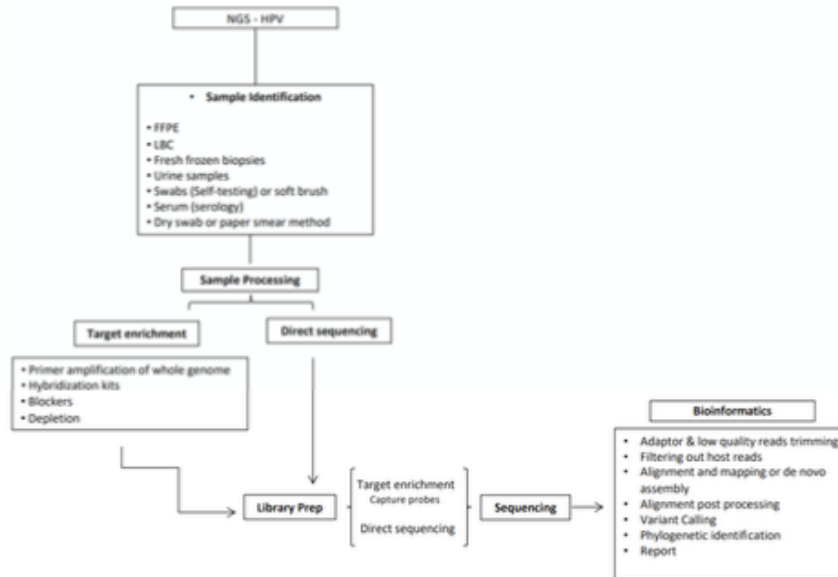
primers/probes which may escape detection using standard molecular approaches [2–4]. When employing whole-genome NGS, which covers the entire genome and not only exons or targeted regions, we conveniently allow high-accuracy determination of sequences below the phylogenetic level of genotype, i.e., variants and subvariants of HPV [3–8]. Indeed, in recent years, various studies have utilized NGS approaches to generate detailed insights into potential disease-related mechanisms of HPV. NGS has shown that certain sublineages of HPV are associated with a higher risk of cancer [9–11], and its high sensitivity also allows the attributable fraction of cervical cancer associated with HPV to be determined with greater precision compared to traditional PCR techniques [12,13]. Further, NGS has identified certain single-nucleotide polymorphisms (SNPs) associated with a higher likelihood of viral persistence [14] and the key role of HPV *E7* gene conservation in cervical cancer development [15].

Although NGS has been used in HPV research for some years with different applications, as described above, NGS in clinical diagnosis is not yet extensively used. Lack of standardization and quality guidelines, as well as expense and requirement for ancillary laboratory infrastructure, may have slowed down the adoption of this technology in clinical laboratories. However, as is discussed later, NGS use could ultimately improve diagnosis and management of patients with HPV-driven lesions. This includes utilities ranging from a more sensitive detection of HPV to detection of true viral persistence [13], identification of risk according to HPV sublineage [9–11], and detection of circulating HPV DNA in patients who have received cancer treatment [16].

NGS is a technology based on massively parallel sequencing or “deep sequencing” of nucleic acid sequences. Nucleic acid sequences are fragmented with each fragment being amplified and sequenced multiple times, providing a depth of information which can deliver accurate data at the nucleotide level. NGS can sequence the entire genome of HPV or be limited to specific areas of interest [11,14].

There are several general approaches to sequencing, depending on the size (i.e., length) of nucleotide reads obtained and detection method employed. Illumina and IonTorrent instruments obtain reads that are approximately 250 base pair (bp) long, whereas Oxford Nanopore Minlon and PacBio obtain longer reads, potentially exceeding 10,000 bp in length [17].

The whole NGS process requires several steps, which include initial sample identification, processing (nucleic acid extraction), viral enrichment (optional), library preparation, sequencing, and bioinformatic analysis of the raw data (Figure 1). While some of these steps are consistent with general requirements for molecular detection (i.e., sample extraction), the described downstream aspects arguably require an additional set of skills and analyses. Given the multistep nature of the process and the generation of large amounts of detailed data generated, it is essential that, where possible, standardization and quality checks to support consistency and integrity of data outputs are considered and implemented. For other applications, including those relevant to bacteriology and molecular pathology, quality guidelines have already been developed [18–20]. However, these are still lacking for the HPV field. Our chief aim is to provide a comprehensive starting point to widen perspectives and give practical advice for those who are new(er) to this topic, thus lowering barriers to introduce NGS specifically in HPV-based research and clinical applications.



**Figure 1.** Next-generation sequencing (NGS) for human papillomavirus (HPV) detection and characterization process steps, from sample preparation to data analysis, with focus on whole-genome sequencing (WGS).

We are active in the field of HPV testing at the International HPV Reference Laboratory and the Scottish HPV Reference Laboratory (SHPVRL). Both entities were established in 2008, under the auspices of WHO (LabNet) and through the National Health Service of Scotland respectively. As such, we are committed to the evaluation and application of new technologies, including NGS technology, specifically to support the prevention and management of human cancers associated with HPV. In the present work we discuss the key stages of HPV-specific applications utilizing NGS and offer practical suggestions for quality-assurance procedures to support these stages, focusing on HPV whole-genome sequencing approaches. We believe that our experience may facilitate the implementation of NGS in laboratory settings so that it can play an increasing role in research, epidemiology, and importantly clinical testing. With respect to the latter, this is an important consideration, given the increased incorporation of NGS systems into routine departments of laboratory medicine, partly because of national strategies designed to harness the benefits of genomic medicine at the patient level in an agile but comprehensive way [21]. NGS is also consistent with the concept and aspirations of precision medicine, defined as an “approach for disease treatment and prevention that takes into account variability in genes, environment and lifestyle for each patient” [22].

**2. Materials and Methods**

*2.1. NGS-Process Step 1: Laboratory Procedures*

*2.1.1. Pre-Analytical Sample Processing*

All specimens used for the identification of HPV by existing molecular tests can in theory be used for NGS. However, various biospecimen types may present their own unique challenges. Liquid-based cytology samples, swab samples, or fresh frozen biopsies may typically be readily applied on an NGS platform. However, cells and tissue derived from formalin-fixed paraffin embedded (FFPE) material are more likely to be associated with fragmented nucleic acid and crosslinks between intracellular macromolecules such as proteins and DNA. Fragmentation can be a rate-limiting factor in approaches that demand



longer amplicons, which is why it is useful to note that researchers have successfully used shorter amplicons when working with FFPE DNA using NGS technologies [23–26].

### 2.1.2. Nucleic Acid Extraction Method

Ideally, extraction methods should be assessed with a pilot panel of samples before embarking on a large project, in order to determine the quality and suitability of the specific extract for downstream NGS. Where possible, we recommend investigators evaluate at least two different extraction technologies to maximize nucleic acid yield. The quality, quantity, and fragment length prior to the library preparation must be determined, as different library preparations may require different nucleic acid input, quality, and length recommendations for library success. For Illumina (San Diego, CA, USA) DNA libraries for example, most of the protocols are optimized for 1 ng of input. Assessing the DNA purity is needed to ensure that the extract does not contain possible contaminants (EDTA, phenol, and ethanol) which can result in assay failure. UV absorbance is a common method used for assessing the purity of a DNA sample, and protocols generally define the “pure”/acceptable range as having an absorbance ratio values of 1.8–2.0 [27,28]. Highly fragmented nucleic acid extraction can lead to missing regions and sections with a low number of reads (low coverage), and the analysis may fail if subjected to enrichment-based amplification methods (see further below) [29,30]. Fragment length can be analyzed using a qPCR with specific length amplicons (same as target amplicon size if PCR-based enrichment is used) or through gel electrophoresis. As described above, a sample being highly fragmented prior to library preparation does not definitively preclude it from sequencing; however, knowing fragment length informs downstream options (e.g., if material is highly fragmented, operators can opt to skip fragmentation steps within the library preparation). HPV positive (with known HPV sequence) and negative internal control samples are necessary to demonstrate that nucleic material has been correctly extracted during the extraction process.

### 2.1.3. Specimen Enrichment Approach

While HPV-positive clinical samples contain HPV nucleic acid, they are naturally dominated by nucleic acid from non-HPV sources (i.e., human and other-microbiome). If NGS is performed directly on the extract, without a targeted approach, the HPV content is at a relatively low proportion vs. the total nucleic acid sample. As the human genome length is 3 billion base pairs vs. ~8000 base pairs for HPV, the relative proportion of human nucleic acids extracted is +200,000%.

Sequencing studies reveal that viruses typically represent less than 1% of the total genomic material detected in a human specimen [7], and therefore, detection of any virus by NGS formerly required subjecting specimens to either (a) host genome depletion or (b) viral enrichment, first. Different approaches to increase the viral component, include low-speed or high-speed gradient centrifugation, separation of long chromosomal DNA, digestion of nucleic acids not protected by virions (e.g., nuclease treatment), filtration to remove bacterial and host cells, or targeted sequence capture [31–36]. Each of these procedures may bias against detection of some viruses; therefore, pilot studies to validate accuracy and reproducibility of the method for the investigator’s specific purpose are necessary. A variety of methods have been described to enhance the HPV content of a sample which are described below.

#### (A). Depletion Protocols Using Saponin- or Lysis-Based Methods

Saponin is a non-ionic surfactant that depletes the human genome affecting the pathogen-human DNA ratio [37,38]. MolYsis (Molzym, Germany) is a commercial product which works through selective lysis of host cells and associated degradation of released host DNA. Both products thus reduce the amount of host nucleic acid, enriching the HPV DNA (or other desired bacterial, fungal, or viral DNA) while simultaneously removing potential PCR inhibitors. However, replicating/intracellular viruses are also depleted with these methods, and potential loss of viral signaling can occur, especially for viruses (such as HPV) that integrate in the human genome.

A particular NGS application related to HPV is the detection of viral integration sites into the human genome. When performing WGS, authors have reported that tumors had either a small or a very large deletion in the viral genome and discovered that these deletions were the result of either HPV integration into the human genome or HPV-HPV sequence junctions [39]. It has further been reported that at least 83% [40,41] of cervical cancers with HPV infection have HPV integration, which can occur at any chromosome but more frequently at certain fragile sites [42]. HPV integration can significantly increase related gene expression and has been associated with a worse survival rate (compared to those with episomal HPV) [43–46]. HPV integration status may therefore have promise as a biomarker for risk stratification [47–49], including the monitoring of treatment and therapy. Different methods have been used to study HPV integration sites (e.g., amplification of HPV oncogene transcripts and detection of integrated HPV sequences by ligation-mediated PCR). With the development of NGS, whole-genome sequencing has been used for virus integration sites detection [50,51]. However, it requires large amounts of sequencing data, and thus is not applicable in clinical usage, which requires fast and accurate results. To date, the development of new NGS methods for HPV integration detection with high accuracy and prompt reporting capacity is ongoing [52]. However, the best way currently to detect integration sites is reached by using probe-captured sequencing methods (see section “Enrichment protocols” below). After enrichment of virus genomic material, the fusion fragment of human and HPV sequence is isolated and further sequenced by NGS.

#### (B). Enrichment Protocols

In addition to depletion protocols, the two most common enrichment protocols for HPV are PCR enrichment in the absence of chemical components and using capture protocols. PCR enrichment involves performing a PCR reaction prior to library preparation. For WGS, this technique may include amplification of the whole genome of the target virus via overlapping primers covering the entire genome (other techniques may not require amplification of the whole genome). A disadvantage of this technique is that there is a need to know the target that is to be sequenced a priori—thus, the method is not valid for detection of novel or nontargeted HPVs. Additionally, in the case of fragmented material, the number of primers required can be high, which requires a structured a priori design approach, and the primers may vary in their affinity for the separate regions. Primer-dimer formation can also limit efficient target amplification, and therefore, pilot studies are needed to validate the protocol for each specific purpose [9,53].

To overcome these bottlenecks, unbiased amplification (not based on PCR) has been commonly used for viral enrichment, as it amplifies all DNA material present in the sample. Multiple displacement amplification (MDA) is the gold standard method for non-PCR-based amplification techniques, where the reaction is based on annealing random hexamer primers to the DNA template [54]. MDA provides an effective way of amplifying minimal quantities of DNA, but there exist biases associated with this technology. Chimera formation, preferential amplification of circular single stranded DNA, and nonuniform amplification of linear genomes have been documented [55,56]. Authors have quantified the amount of amplification of both human DNA and HPV DNA by adding 20 copies/ $\mu\text{L}$  of HPV 16 plasmid to samples of human placental DNA at 1 ng/ $\mu\text{L}$  and reported an amplification of 26-fold for human DNA and 679-fold for HPV 16 DNA, suggesting that MDA is a good method for enriching circular HPV genomes [5]. Using MDA and NGS sequencing, researchers have been able to detect a plethora of novel HPVs as well as known HPV types, not detected by traditional PCR-based enrichment methods, among skin lesions/tumors and condyloma accuminata [3,4,7].

Another enrichment approach is to use a set of specific probes, “baits”, to recover HPV sequences from the entire genome of the virus. In brief, labelled biotinylated HPV specific probes are captured by streptavidin coated magnetic beads after hybridization, resulting in “pure” HPV-derived reads. Consistent with the overlapping PCR approach, these probes need to be designed in advance but can be modified as required if low read numbers are obtained for some regions of the viral genome. Although it can be expensive,



the advantage of the probes/baits approach is the large number of different probes one may include, allowing thousands of probes in one design. This means that it is possible to load the analysis with probes for all known HPVs, depending on purpose. Furthermore, it is the current gold standard method for integration studies [52].

Regardless of protocol applied, to validate the quality of host depletion or viral enrichment, positive and negative control material must be added at the nucleic acid extraction step and carried through the enrichment/capture/depletion step and all stages of subsequent analysis. As a positive control, cell line material infected with HPV is commonly used. However, we would recommend using specimens that contain both human DNA and HPV DNA in the typical concentrations that would correspond to real-life clinical specimens (99:1), as the sensitivity of the different approaches may vary. As a negative control, human DNA (HPV free, but containing the corresponding background “noise” as a true HPV-negative sample) or DNA-free water can be used. Note that if primer-based target enrichment is used, negative controls may contain noise such as primer-dimer bands when assessed by electrophoreses as well as unspecific amplification; these should be clearly discriminated from target sequences. Again, the introduction of homogeneous internal quality controls (IQCs) in the nucleic acid amplification should help in the identification of such issues. Internal positive control material demonstrates whether the depletion/lysis has removed/reduced the HPV target in large proportion during the sample preparation stage. For the enrichment step, positive control material should reflect if amplification of the target regions/genome occurs and negative controls help determine whether any nonexpected amplification/targeting occurred.

#### 2.1.4. Direct Sequencing

Several operators have opted for performing WGS directly after nucleic acid extraction, without enrichment. As described above, reaching a high sequencing depth is required, due to the low proportion of viral sequences typically present in the human specimens. Direct sequencing has enabled an agnostic approach to DNA presence in clinical samples: while around 10% of cervical cancers are found to be negative for oncogenic HPVs by traditional PCR-based genotyping methods [57–61]; direct sequencing can reveal the presence of viral sequences of potentially causative HPVs in such cases [13,62]. Interestingly, most of the HPVs detected with direct sequencing among the carcinomas negative by traditional PCR typing systems corresponded to HPV types within the explicit detection range of these assays. Therefore, the information obtained by direct sequencing can be used, not only to detect novel variants or types that may have escaped traditional amplification but also as a way to quantify and monitor shortfalls in detection due to sensitivity issues. Furthermore, evidence suggests that HPV-negative cancer patients have a worse longitudinal prognosis [60,63,64]; this is reflected in the staging system for oropharyngeal cancer which acknowledges the dichotomous disease status based on HPV presence [65] as does the recent WHO update for female genital tumors [66]. How “best” to annotate HPV status in cancer tissue is an area which arguably lacks consensus in the literature; however, NGS provides a powerful tool to at least resolve which cancer cases may be truly virally negative.

#### 2.1.5. Library Preparation and Sequencing

At present, there are several different sequencing chemistries available. Each system has its own protocols and due to the diversity of platforms, and rapid pace of developments, we cannot recommend a specific one for all HPV applications. Akin to the assessment and introduction of any new technology in-house, we strongly recommend validation that includes confirmation of expected results from “known” quality materials and the evaluation of different kits with the specific analytical purpose in mind. For laboratories looking to embed NGS into the accredited scope of their clinical service, initial validation followed by yearly verification would likely be mandatory.

Quality, quantity, and fragment length analysis is a must to confirm success of library preparation. Library preparation protocols usually inform the operator about the concentration and size expected for prepared libraries. If measurements do not reach the expected values, (e.g., fragments are too big/small or the library concentration is too diluted) optimization of fragmentation times, clean-up processes, or amplification steps should be performed. Larger fragments cluster less efficiently than smaller molecules, and a low concentration of prepared libraries translates into a low number of sequenced reads. Here, it is crucial to consider in which context libraries are analyzed; in a research study, suboptimal measurements can occasionally be acceptable but probably never in a clinical context where protocol adherence is paramount and actionable test results are needed.

One also needs to perform accurate normalization to obtain homogeneous distribution (number of sequencing reads) of the samples and assure that the proper sequencing read length is used, depending on the insert size of the library. As an example, Illumina libraries prepared with dual-indexing that show a fragment length of 200 bp should not be sequenced with  $2 \times 150$  bp, as part of the fragment length (around 130 bp) corresponds to adapter sequences, and the insert size, which is the actual query sequence, only comprises 70 bp ( $200 - 130$  bp). Thus, 80 bp of the 150 bp sequenced ( $150 - 70$  bp) does not contribute useful information. Libraries from positive and negative samples must be added into the final input dilution. Our recommendations for quality-control steps in the NGS workflow are summarized in Table 1.

**Table 1.** Steps, potential quality issues, and proposed mitigations for next-generation sequencing (NGS) analytical workflow, with a focus on whole-genome sequencing purposes.

Human Papillomavirus Detection by Next-Generation Sequencing		
NGS Step	Possible Difficulties	Mitigations
Sample preparation	Nucleic acid quality and/or quantity outside library prep kit requirements	Selection of appropriate nucleic acid extraction methods. Pilot study comparing different extraction kits. Selection of enrichment or depletion protocol.
	Incorrect fragment size (too short or too long)	Introduction of homogeneous internal quality controls. Electrophoresis, bioanalyzer, and/or fluorometric quantitation. Selection of further protocols based on the fragment size (e.g., use of shorter amplicons if DNA is highly fragmented)
Library preparation and sequencing	Incorrect fragment size (too short or too long)	Correct selection of library kit and fragment length.
	Incorrect number of sequencing reads or partial reads	Correct selection of sequencing kit (e.g., 75 bp and 150 bp) to avoid sequencing adapters or longer fragments that insert size.
Data analysis	Low sequencing depth	Library preparation and sequencing piloting, and re-analysis Confirm reference sequence is correct. Use of updated database.
	Incorrect alignment	In case of low sequencing depth at the beginning or end of the reference sequence, note that HPV is circular and not linear as the reference. Confirmation that desired alignment cut-offs are correct.
	Mix/chimeras of microbial organisms	Filter reads—use of updated databases and careful settings of parameters. De novo assembly evaluation (HPV Chimera scripts)
Storage	Validation of pipeline	Digital IQC, EQA, external assessment. Interlaboratory comparison
	Large amount of data	Cloud services, compression of files, and storage of only raw input and final output.
	Security	Restricted super-user access, individually curated data access Analyst working only with coded/pseudonymized samples (where an independent database administrator holds the key code at another site)
	Length of data storage	Organization policy/data archiving laws and regulations

## 2.2. NGS-Process Step 2: Bioinformatical Analysis

### 2.2.1. Raw Sequence Data Management

The output data from a sequencing machine are often referred to as raw data. Raw data management generally includes filtering steps to remove poor-quality data and host-derived human reads and continues by mapping nonhuman high-quality reads directly to a

known reference database or performing a de novo assembly approach, finishing with HPV taxonomy classification, phylogenetic analysis, and variant calling. There are several open-access tools that can be used to analyze “big data”. A set of bioinformatic algorithms, when executed in a predefined sequence, is collectively referred to as a bioinformatics “pipeline”. These pipelines can be designed in-house by teams with available bioinformatics expertise or obtained as ready-to-use applications from commercial suppliers. The premade pipelines are principally aimed to users with little bioinformatic experience and act as a “blackbox” (user does not know which algorithm and calculation(s) are used by the pipeline).

### 2.2.2. Sequence Analyses: Quality Assessment of Reads

Each bioinformatic tool deployed in the pipeline uses different algorithms and parameters when handling data. When the objective of a project is to resolve a 0.5–1% difference between sequences, these differences in parameter settings could mean a different interpretation on nucleotide/mutation level is reached depending on which pipeline is being used. Ergo, two studies on the same nucleotide position could reach two different conclusions as to whether a mutation is present.

Errors due to poor sample handling and storage conditions, polymerase bias, PCR- or qPCR-induced errors, and incorporation errors within sequencing may be introduced during sample preparation, amplification, library preparation, and sequencing stages. While these errors might not interfere with the identification of an HPV type (where the sequence divergence is 10% relative to its most closely related type), they might compromise the identification of sublineages or variations in nucleotides that could have implications on accuracy, consistency, and, potentially, predicted phenotype. Careful sample handling, selection of a high-fidelity polymerase [67], and quality control of the raw data is a must [68].

Most analysis applications use FASTQ files as input for analysis; however, different sequencing instruments may give different extensions for raw sequencing data (e.g., Illumina generates bcl files), and the first step is the conversion of those files to a standard format (FASTQ). Platforms usually provide software for the desired conversion (e.g., bcl2fastq from Illumina). Raw FASTQ files should be subjected to quality trimming and adaptor removal as a first step. Software for quality trimming and adaptor removal include Cutadapt, Trimmomatic, Trim Galore!, SeqTrim, and FastX among others [69]. Quality trimming is performed to remove low quality reads and aims to reduce the effect of the progressive decrease in sequencing quality with the increased length of the sequenced library. Trimming removes low quality portions of NGS reads while preserving the high-quality part of such a read. The user can specify the quality cut-off for a base or use a “sliding window” approach (defined as setting a cut-off for the average quality detected in a number of X contiguous bases instead of just one base). Quality is usually checked according to the Phred quality scores, which are scores logarithmically related to base-calling error probabilities [70]. As an example, a Phred quality score of Q30 corresponds to a base calling accuracy of 99.9% (1 error per 1000 bp). The minimum quality recommended in the literature is a Phred quality score of 20 (99.0% accuracy; 1 error per 100 bp), with the optimal quality however being above Q30 [71].

### 2.2.3. Alignment of Reads to a Suitable Human Reference Genome

To obtain a dataset that contains only reads of interest, e.g., viral-related reads for HPV detection, nontarget sequences may be filtered out at the bioinformatics level to speed up downstream analysis and decrease the risk of misassemblies of genomic data. Most researchers applying WGS from sequenced extracted material opt for filtering out human genome sequences only (leaving HPV plus “other” microorganism reads). This is practical as human reads (according to our experience of multiple sequencing projects with a metagenomic perspective) account for approximately 90% of the total (with some variability depending on the sample origin) [7,13,62]. While several successive “versions” corresponding to human genome reference exist, it is recommended that the latest build



(released in December of 2013), officially named GRCh38 (Genome Research Consortium human build 38) or commonly Hg38 (human genome build 38) is used.

Numerous aligners exist so far and are being developed in order to achieve greater accuracy pertaining to precision. Widely used tools include BWA-MEM (Burrows-Wheeler aligner) [72], SOAP2 [73], or Bowtie 2 [74], but several commercial tools are also available such as NextGenmap and Novoalign [75,76].

While some operators may, at this point, choose to filter out reads that are identical to the human reference genome (100% identical), another approach is to employ looser parameters which accept reads as human if the identity and coverage across the human genome sequence of interest are at least 95% and 75%, respectively. This latter approach allows for the detection of possible mutations not anticipated in the reference sequence, although this flexibility should be tempered so as not to misclassify nonhuman reads as human. An important thing to consider after performing the aligning/mapping to the human reference genome is to select which “unmapped” reads are to be used in downstream analysis. Sequencing with paired-end reads may contain (1) paired-end reads where both reads are unmapped, (2) paired-end reads where one of the pair read maps to the genome and the other does not. Operators may decide to discard nonhuman single reads and continue only with nonhuman paired-end reads, or to include them all; such a choice is entirely dependent on coverage obtained and the aim of the project/resolution required.

#### 2.2.4. Alignment of Reads to a Suitable HPV Reference Database

Once human (or host) sequences are filtered out from the high-quality data set, most operators align reads to a known and curated HPV database for HPV classification or to the reference HPV genome in question. Currently, there are 222 different HPV types officially established (data accessed on 18 March 2021 from Hpvcenter.se) and another 220 putative novel HPV types (not cloned and investigated by the International HPV Reference Center) whose complete sequence can be found in the public database from the papillomavirus Episteme (data accessed on 18 March 2021 at <https://pave.niaid.nih.gov/>).

Furthermore, there are many partial genomic sequences of HPV isolates (not all specifying which HPV type they correspond to) available at public databases, with GenBank having >33,500 hits retrieved when typing “human papillomavirus” (data accessed on 18 March 2021). A recent publication [77] detected up to 0.5% chimeric sequences and/or taxonomy errors when analyzing HPV sequences in the GenBank database. This highlights the importance that the database be obtained from a quality reference repository and that local curation is performed before doing any type of alignment, such as checking for potential errors and updates. In-house databases belonging to individual investigators should be periodically updated with canonical or reference types, as new HPV types are continuously being discovered [5,78].

It is particularly key to note that, when aligning the sequencing reads to a specific HPV genotype, operators ascertain that the correct reference sequence is used. There are 3650 sequences belonging to HPV16 isolates in GenBank (sequence length 7500–8500, data accessed on 18 March 2021), showing differences that may reach up to 10% of the total genome. If each separate investigation were to use a different sequence as reference genome, then comparison between publications becomes challenging at best. The reference genomes that should be used for each HPV type are provided at the International HPV Reference Laboratory website (Hpvcenter.se); accessed 7 July 2021, as well as at the papillomavirus Episteme database (<https://pave.niaid.nih.gov/>; accessed 7 July 2021). The latter resource has a contemporary collection of internationally ratified sequences from the reference clones corrected for known sequencing mistakes in the original sequences.

#### 2.2.5. Identification of HPV Types/Lineage/Sublineages

Classification of HPVs is based on the nucleotide sequence homology of the *L1* gene, which is the most conserved region of the viral genome. Within the family, different

genera share less than 60% nucleotide similarity. Within each genus, different species share between 60% and 70% similarity. Below the species level, a novel HPV type shares less than 90% similarity to any other type [79–81]. The definition of a variant lineage is that the L1 open-reading frame differs by more than 1%, but less than the 10% that would make it another HPV type [82]. A variant sublineage is defined as groups of sequences with 0.5–1.0% differences between genomes [83].

There exist different tools for the identification of variants. One of the most used and user friendly is BLAST [84]. This tool compares the sequence under investigation to sequences stored in the database, detailing statistical significance of matches. Again, the importance of using standard references for HPV variants is essential. Burk et al. described the representative genomes for viral variant lineages and sublineages, and most authors rely on these sequences as variant lineages references [82].

If a phylogenetic analysis is required, different open-source tools exist (RaxML, MegaX) that infer phylogenetic trees after choosing the statistical method [85,86].

#### 2.2.6. Evaluation of Coverage across the Genome

If the purpose is to detect HPV genotypes (not within-genotype specific variant calling), once the reads are aligned to the HPV database, we recommended that cut-offs are applied on which HPV positivity is based (Table 1).

This could be, e.g., setting a minimum of 10 reads detected for a specific HPV type together with a coverage of at least 10% of the HPV genome (around 800 bp coverage). This approach would avoid false positivity generated by background noise (e.g., presence of many low complexity reads mapping to just a small region of the genome). If phylogenetic analysis or variant calling is required, a FASTA file with the “query” sequence must be created. When creating a FASTA file from the obtained sequencing reads/contigs, investigators should be aware of the extent of genome coverage to see if there are missing regions

Evaluating the full coverage of the sequence is important, as several tools that convert the sequencing reads into a FASTA file use a reference sequence to account for the regions that are not covered. The use of “N”s is recommended for the positions that are not covered by the sequencing reads. For variant lineage assignment, exclusion must be considered for specimens with poor read depth (<200 median depth) and/or low genome coverage (<80% genome coverage) [53]. For variant calling, even stricter cutoffs should be applied. Additionally, further steps including marking duplicates to identify read pairs likely to have originated from duplicates of the same original DNA fragments and recalibration of base quality scores should be performed, as suggested by the best practices at GATK [87]. Considering just the base depth as a cutoff for variant calling (e.g., five reads per position) is not enough to assure accurate calling. It is essential to differentiate between true positive variants and false positive variants. Parameters and statistics which describe how many reads cover the variant, what proportion of reads are in forward vs. reverse orientation, and what the sequence context is like around the variant site should be considered.

#### 2.2.7. De Novo Assembly of HPV Contigs

When the correct reference genome is not known, the (re-)construction of the sequenced genome must be performed without a priori knowledge of either the correct original sequence (or the order of the DNA fragments), by assembling overlapping reads into one or more contigs. This process is known as de novo assembly. Subsequent post assembly assessment is mandatory to reduce the risk of chimeric sequences and possible miscalling of HPV positivity in samples and/or erroneous calling of new HPV variants/genotypes. HPV-Chimera scripts exist to help researchers determine the accuracy of their HPV contigs [12,77].

#### 2.2.8. Digital Quality Assessment

While positive and negative controls can be incorporated into laboratory experiments and several quality check-points are available during the whole laboratory process, we cur-

rently lack an agreed approach for the quality control of bioinformatical tools and pipelines. Digital IQCS can be prepared from confirmed and verified positive material and stored as FASTQ files (raw data). Interlaboratory exchange of data can provide reassurance by comparing results on sequences derived from the same specimens. However, this requires resources and collaboration which may not always be available in the short term. Therefore, it would be beneficial to have positive IQCS available in an online repository that could be used to verify the pipeline when setting up a new service/test or when a tool is updated.

#### 2.2.9. Journal Submission Requirements

Recently, several journals have started to request that all authors who submit manuscripts containing NGS data provide a detailed summary of sequencing coverage and quality statistics. For example, the International Journal of Cancer requires a summary from submitting authors that must include all information about library preparation, sequencing technology information (e.g., platform, read length, and paired-end/single read approach), as well as preprocessing, quality control, and filtering of the raw NGS data [88]. Furthermore, the sequencing coverage and quality statistics of each sample must be summarized as a Supplementary Table.

#### 2.2.10. HPV NGS in Clinical Settings

Application of NGS for clinical testing requires a level of quality assurance and monitoring likely to be even more stringent than systems set up in research laboratories. Verification and validation for each of the steps that make up the NGS process is the key to obtain and provide reliable results to clinicians and patients. Any minor change of the wet-lab protocol or any parameter in the data analysis requires a full verification with previously known samples/sequences. This in combination with the external quality assessment, and accreditation helps ensure the validity of the clinical results.

#### 2.2.11. External Quality Assessment and Accreditation

Suppliers of EQA schemes have developed external quality materials to support NGS sequencing results for various pathogens and human genes (e.g., GenQA, Statens Serum Institut, and QCMD (pilot)). At time of publication, we are not aware of any official HPV EQA scheme to support NGS for wet and in silico analysis or indeed data/dry analysis. This is arguably a current deficit, as while interlaboratory exchange of materials is undoubtedly helpful for quality assurance and validation of HPV NGS, such exchanges do not wholly “stand in” for consistent performance in a formal accredited EQA scheme(s). Should HPV NGS move to the diagnostic context, then this would increasingly require address.

#### 2.2.12. Data Storage Requirements

Data storage demands of NGS are often very large and need to be carefully considered before local implementation. Unfortunately, there is no (international) consensus on what and how data should be stored beyond the general recommendation that it should be stored in line with national and local capacity and governance policies. In high-level terms, we recommend storing the raw FASTQ files in a compressed mode (.fastq.gz), the final output, and the full-log file (documentation on the software/tools used, including versions, parameters, and github location, needed to obtain the output files) making the whole analysis reproducible.

If intermediate files are to be kept, they should be stored as standard open-file formats FASTQ, BAM, and VCF, facilitating the exchange with other laboratories, where governance permits. Cloud-based storage could be a very helpful tool; however, this may be challenging to reconcile with data protection. Currently, most journals ask for data availability and request that authors upload all nonhuman sequences detected in the study to different databases (such as the European Nucleotide Archive (ENA), Sequencing Read Archive (SRA), and Genbank) to make research publicly available and re-usable for other scientists without compromising confidentiality.



### 3. Discussion

Next-generation sequencing (NGS) has enabled researchers to detect human papillomavirus (HPV) infections with unprecedented sensitivity and accuracy while simultaneously providing the whole viral sequence to be analyzed for viral point mutations, variant lineages, and genome variations. Increasing incorporation of NGS into laboratories for research, epidemiology, and diagnostic purposes may be only a matter of time, particularly as costs reduce with increasing demand and competition.

Certainly, the use of NGS for clinical workstreams generally requires accreditation/auditing from an independent regulator. As an example, the National Accreditation Body for the United Kingdom (UKAS) now has expertise to assess and accredit labs that have NGS in scope, working to ISO 15189:2012 standards. This process covers assessment of staff training, quality control of pipelines, initial validation of the whole process compared to a gold-standard approach, and reproducibility of results plus checks to ensure security of data and access is correct. The magnitude of the effort to achieve accreditation in a particular service laboratory is likely to depend on local support for the integration of NGS to support public health epidemiology and precision medicine in general terms. The SARS-CoV-2 pandemic has brought about an exponential increase in molecular testing and associated infrastructure, including that required to support sequence-level variant detection. This is likely to pay dividends for the establishment of “cross organism working” to a diagnostic standard [89]. Certainly, NGS has the potential to support risk stratification of patients with HPV-associated disease through its ability to detect types within tissue with exquisite sensitivity and through providing subtle insights into aspects of HPV infection that may be predictive of outcome (integration pattern and status, delineation of dominant variant(s), and viral load). NGS may also be applied to liquid biopsies, from blood, to support longitudinal monitoring of treatment success [90].

The bulk of HPV sequencing to date has been performed using the IonTorrent initially [9,11,91], followed by the newer Illumina platform [7,53,57]; some very recent studies have also used nanopore platforms [52,92]. While many studies report higher sensitivity and accuracy when comparing NGS to routine genotyping methods [6,8], there are very few studies where one HPV NGS approach has been directly compared to another in a head-to-head approach using the same sample set [93,94]. Hence, there are fewer data available compared to the relative wealth of peer-reviewed literature on head-to-head performance of traditional HPV molecular assays. While many researchers are already using NGS for detection and analysis of HPV-associated diseases, there is no clear international consensus about what steps are to be performed nor which quality criteria are appropriate.

Typically, published presequencing laboratory protocols appear well validated for providing detail on clear-quality assessment procedures and specifying the requirements needed for library preparation evaluation and success. Nevertheless, even though the inclusion of positive and negative controls should be a must, the presence and/or description of these controls is not commonly found within NGS publications. This may be in part due to the relatively high cost of NGS. Hornung et al. reviewed ~265 publications which had used NGS for microbiome research and found that only 30% of publications used any type of negative controls and that less than 10% reported positive controls. Additionally, they observed that some of the results reported were potentially indistinguishable from contaminants [95].

Furthermore, to the best of our knowledge, quality assessment of bioinformatic analysis appears not to be as standardized, when compared to the analytical sample handling stages. However, we consider it key that the use of a curated and updated databases, use of standard reference genome sequences, and description of the parameters used for each step should be described fully in all publications.

The present piece aims to describe and summarize the application of WGS for HPV detection and has mainly focused on DNA detection to detect the whole HPV sequence. Nevertheless, RNA sequencing should not be forgotten as an alternative method for HPV detection, as RNA transcription not only enables detection of HPVs but provides further

information on the active HPV infection that drives viral oncogene expression. In the cancers that are known to be caused by HPV, transcription of viral genes is necessary for viral pathogenicity [96]. Studying transcription of the E6 and E7 oncogenes has been useful to elucidate which infections are likely to be involved in the etiology of the tumor, e.g., in head and neck cancer studies [97–99]. Even though RNA sequencing is not usually performed within studies aiming to analyze the whole HPV genome (as noncoding regions, e.g., URR, are not detected if RNA extraction and sequencing are performed), important information about active infection and viral oncogene expression is obtained with this approach.

International collaboration is essential to efficiently further knowledge, scientific development, and concerted efforts to combat globally prevalent viral infections. In the case of HPV, the International HPV Laboratory Network LabNet was created by WHO in 2006 to support global development of laboratory standardization and quality assurance of HPV detection methods. LabNet concentrated on evaluating and improving methods used for research and evaluation of HPV-based screening and vaccination [100]. As part of this effort, LabNet published an HPV laboratory manual, based on knowledge and experience gained through international collaborative studies, aiming to assist in establishing the laboratory support required for HPV research [101]. The successor to LabNet, the International HPV Reference Center (Hpvcenter.se), aims to support reliable and comparable HPV detection services, allowing data to be internationally comparable. The Center has organized and issued global proficiency panels (sets of blinded samples containing HPV genotypes at different concentrations) since 2008, and a definite improvement in average assay performance globally has been seen since the panel was issued [102]. The SHPVRL has also acted as a hub laboratory to support the creation of materials and best practice documents to facilitate the introduction of new HPV technologies and their continued monitoring [103,104].

Together, we now work between our two laboratories to exchange samples, know-how, and protocols for bioinformatical flow and sequence analyses. Hopefully, this will strengthen the quality of work produced by both settings and act as a catalyst/model for future international endeavors. We propose that it is time for NGS to be included in the global efforts on quality assurance and improvement in HPV-based testing and diagnostics. By establishing a set of quality standards and best-practice statements, the community could systematically develop and apply NGS guidelines suited for HPV research, epidemiology, and diagnostics to ensure this innovative and powerful technology is developed in an internationally comparable and robust manner.

**Author Contributions:** Conceptualization, K.C. and K.S.; methodology, L.S.A.M. and D.G.; writing—original draft preparation, L.S.A.M. and D.G.; writing—Review and Editing, K.C. and K.S.; supervision, K.C. and K.S.; funding acquisition, L.S.A.M. and K.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Center for Innovative Medicine (CIMED grant number 613/06, to KS), the Swedish Medical Society (SLS, grant number 885941, to KS), and the Swedish Foundation for Strategic Research (grant number RB13-0011, supporting KS and SAM). The SHPVRL is supported by National Services Division of the National Health Service in Scotland.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest related to this work.

## References

1. Lorincz, A.; Wheeler, C.M.; Cuschieri, K.; Geraets, D.; Meijer, C.J.L.M.; Quint, W. Developing and Standardizing Human Papillomavirus Tests. In *Human Papillomavirus: Proving and Using a Viral Cause for Cancer*; David Jenkins, F.X.B., Ed.; Academic Press: Cambridge, MA, USA, 2020; pp. 111–130.



2. Gradissimo, A.; Burk, R.D. Molecular tests potentially improving HPV screening and genotyping for cervical cancer prevention. *Expert Rev. Mol. Diagn.* **2017**, *17*, 379–391. [CrossRef]
3. Arroyo Muhr, L.S.; Bzhalava, D.; Lagheden, C.; Eklund, C.; Johansson, H.; Forslund, O.; Dillner, J.; Hultin, E. Does human papillomavirus-negative condylomata exist? *Virology* **2015**, *485*, 283–288. [CrossRef]
4. Arroyo Muhr, L.S.; Hultin, E.; Bzhalava, D.; Eklund, C.; Lagheden, C.; Ekstrom, J.; Johansson, H.; Forslund, O.; Dillner, J. Human papillomavirus type 197 is commonly present in skin tumors. *Int. J. Cancer* **2015**, *136*, 2546–2555. [CrossRef]
5. Bzhalava, D.; Muhr, L.S.; Lagheden, C.; Ekstrom, J.; Forslund, O.; Dillner, J.; Hultin, E. Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci. Rep.* **2014**, *4*, 5807. [CrossRef]
6. Arroyo, L.S.; Smelov, V.; Bzhalava, D.; Eklund, C.; Hultin, E.; Dillner, J. Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **2013**, *58*, 437–442. [CrossRef] [PubMed]
7. Bzhalava, D.; Johansson, H.; Ekstrom, J.; Faust, H.; Moller, B.; Eklund, C.; Nordin, P.; Stenquist, B.; Paoli, J.; Persson, B.; et al. Unbiased approach for virus detection in skin lesions. *PLoS ONE* **2013**, *8*, e65953. [CrossRef]
8. Nilyanimit, P.; Chansaenroj, J.; Poomipak, W.; Praianantathavorn, K.; Payungporn, S.; Poovorawan, Y. Comparison of Four Human Papillomavirus Genotyping Methods: Next-generation Sequencing, INNO-LiPA, Electrochemical DNA Chip, and Nested-PCR. *Ann. Lab. Med.* **2018**, *38*, 139–146. [CrossRef] [PubMed]
9. Cullen, M.; Boland, J.F.; Schiffman, M.; Zhang, X.; Wentzensen, N.; Yang, Q.; Chen, Z.; Yu, K.; Mitchell, J.; Roberson, D.; et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* **2015**, *1*, 3–11. [CrossRef] [PubMed]
10. Clifford, G.M.; Tenet, V.; Georges, D.; Alemany, L.; Pavon, M.A.; Chen, Z.; Yeager, M.; Cullen, M.; Boland, J.F.; Bass, S.; et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res.* **2019**, *7*, 67–74. [CrossRef]
11. Mirabello, L.; Yeager, M.; Cullen, M.; Boland, J.F.; Chen, Z.; Wentzensen, N.; Zhang, X.; Yu, K.; Yang, Q.; Mitchell, J.; et al. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **2016**, *108*. [CrossRef] [PubMed]
12. Arroyo Muhr, L.S.; Lagheden, C.; Hassan, S.S.; Kleppe, S.N.; Hultin, E.; Dillner, J. De novo sequence assembly requires bioinformatic checking of chimeric sequences. *PLoS ONE* **2020**, *15*, e0237455. [CrossRef]
13. Arroyo Muhr, L.S.; Lagheden, C.; Lei, J.; Eklund, C.; Nordqvist Kleppe, S.; Sparen, P.; Sundstrom, K.; Dillner, J. Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR. *Br. J. Cancer* **2020**, *123*, 1790–1795. [CrossRef]
14. Perez, S.; Cid, A.; Araujo, A.; Lamas, M.J.; Saran, M.T.; Alvarez, M.J.; Lopez-Miragaya, I.; Gonzalez, S.; Torres, J.; Melon, S. A novel real-time genotyping assay for detection of the E6-350G HPV 16 variant. *J. Virol. Methods* **2011**, *173*, 357–363. [CrossRef] [PubMed]
15. Mirabello, L.; Yeager, M.; Yu, K.; Clifford, G.M.; Xiao, Y.; Zhu, B.; Cullen, M.; Boland, J.F.; Wentzensen, N.; Nelson, C.W.; et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* **2017**, *170*, 1164–1174.e6. [CrossRef]
16. Lee, J.Y.; Cutts, R.J.; White, I.; Augustin, Y.; Garcia-Murillas, I.; Fenwick, K.; Matthews, N.; Turner, N.C.; Harrington, K.; Gilbert, D.C.; et al. Next Generation Sequencing Assay for Detection of Circulating HPV DNA (cHPV-DNA) in Patients Undergoing Radical (Chemo)Radiotherapy in Anal Squamous Cell Carcinoma (ASCC). *Front. Oncol.* **2020**, *10*, 505. [CrossRef]
17. Besser, J.; Carleton, H.A.; Gerner-Smidt, P.; Lindsey, R.L.; Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* **2018**, *24*, 335–341. [CrossRef]
18. Gargis, A.S.; Kalman, L.; Lubin, I.M. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *J. Clin. Microbiol.* **2016**, *54*, 2857–2865. [CrossRef] [PubMed]
19. Lopez-Labrador, F.X.; Brown, J.R.; Fischer, N.; Harvala, H.; Van Boheemen, S.; Cinek, O.; Sayiner, A.; Madsen, T.V.; Auvinen, E.; Kufner, V.; et al. Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *J. Clin. Virol.* **2021**, *134*, 104691. [CrossRef] [PubMed]
20. Endrullat, C.; Glokler, J.; Franke, P.; Frohme, M. Standardization and quality management in next-generation sequencing. *Appl. Transl. Genom.* **2016**, *10*, 2–9. [CrossRef]
21. Scottish Science Advisory Council. Informing the Future of Genomic Medicine in Scotland. Available online: <https://www.scottishscience.org.uk/sites/default/files/article-attachments/Genomics%20Full%20Report.pdf> (accessed on 26 May 2021).
22. Medlineplus. Available online: [Medlineplus.gov/genetics/understanding/precisionmedicine/definition/](https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/) (accessed on 26 May 2021).
23. Wong, S.Q.; Li, J.; Tan, A.Y.; Vedururu, R.; Pang, J.M.; Do, H.; Ellul, J.; Doig, K.; Bell, A.; MacArthur, G.A.; et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med. Genom.* **2014**, *7*, 23. [CrossRef]
24. Yost, S.E.; Smith, E.N.; Schwab, R.B.; Bao, L.; Jung, H.; Wang, X.; Voest, E.; Pierce, J.P.; Messer, K.; Parker, B.A.; et al. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **2012**, *40*, e107. [CrossRef] [PubMed]
25. Kerick, M.; Isau, M.; Timmermann, B.; Sultmann, H.; Herwig, R.; Krobitsch, S.; Schaefer, G.; Verdorfer, I.; Bartsch, G.; Klocker, H.; et al. Targeted high throughput sequencing in clinical cancer settings: Formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genom.* **2011**, *4*, 68. [CrossRef]

26. Graw, S.; Meier, R.; Minn, K.; Bloomer, C.; Godwin, A.K.; Fridley, B.; Vlad, A.; Beyerlein, P.; Chien, J. Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci. Rep.* **2015**, *5*, 12335. [CrossRef] [PubMed]
27. Nanodrop. Technical Support Bulletin. Available online: [https://bio.davidson.edu/projects/gcat/protocols/NanoDrop\\_tip.pdf](https://bio.davidson.edu/projects/gcat/protocols/NanoDrop_tip.pdf) (accessed on 26 May 2021).
28. Illumina. Nextera®DNA Library Prep Reference Guide. Available online: [https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\\_documentation/samplepreps\\_nextera/nextera-dna-library-prep-reference-guide-15027987-01.pdf](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera-dna-library-prep-reference-guide-15027987-01.pdf) (accessed on 26 May 2021).
29. Do, H.; Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clin. Chem.* **2015**, *61*, 64–71. [CrossRef]
30. Bettoni, F.; Koyama, F.C.; De Avelar Carpinetti, P.; Galante, P.A.F.; Camargo, A.A.; Asprino, P.F. A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology. *Exp. Mol. Pathol.* **2017**, *103*, 294–299. [CrossRef]
31. Duncavage, E.J.; Magrini, V.; Becker, N.; Armstrong, J.R.; Demeter, R.T.; Wylie, T.; Abel, H.J.; Pfeifer, J.D. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J. Mol. Diagn.* **2011**, *13*, 325–333. [CrossRef]
32. Allander, T.; Emerson, S.U.; Engle, R.E.; Purcell, R.H.; Bukh, J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 11609–11614. [CrossRef]
33. Duhaime, M.B.; Sullivan, M.B. Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **2012**, *434*, 181–186. [CrossRef]
34. Depledge, D.P.; Palser, A.L.; Watson, S.J.; Lai, I.Y.; Gray, E.R.; Grant, P.; Kanda, R.K.; Leproust, E.; Kellam, P.; Breuer, J. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* **2011**, *6*, e27805. [CrossRef] [PubMed]
35. Koehler, J.W.; Hall, A.T.; Rolfe, P.A.; Honko, A.N.; Palacios, G.F.; Fair, J.N.; Muyembe, J.J.; Mulembekani, P.; Schoepp, R.J.; Adesokan, A.; et al. Development and evaluation of a panel of filovirus sequence capture probes for pathogen detection by next-generation sequencing. *PLoS ONE* **2014**, *9*, e107007. [CrossRef]
36. Wylie, T.N.; Wylie, K.M.; Herter, B.N.; Storch, G.A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **2015**, *25*, 1910–1920. [CrossRef]
37. Ji, X.C.; Zhou, L.F.; Li, C.Y.; Shi, Y.J.; Wu, M.L.; Zhang, Y.; Fei, X.F.; Zhao, G. Reduction of Human DNA Contamination in Clinical Cerebrospinal Fluid Specimens Improves the Sensitivity of Metagenomic Next-Generation Sequencing. *J. Mol. Neurosci.* **2020**, *70*, 659–666. [CrossRef] [PubMed]
38. Hasan, M.R.; Rawat, A.; Tang, P.; Jithesh, P.V.; Thomas, E.; Tan, R.; Tilley, P. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *J. Clin. Microbiol.* **2016**, *54*, 919–927. [CrossRef]
39. Gao, G.; Wang, J.; Kasperbauer, J.L.; Tombers, N.M.; Teng, F.; Gou, H.; Zhao, Y.; Bao, Z.; Smith, D.I. Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. *BMC Cancer* **2019**, *19*, 352. [CrossRef] [PubMed]
40. Wentzensen, N.; Vinokurova, S.; Von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **2004**, *64*, 3878–3884. [CrossRef]
41. Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **2017**, *543*, 378–384. [CrossRef] [PubMed]
42. Chandrani, P.; Kulkarni, V.; Iyer, P.; Upadhyay, P.; Chaubal, R.; Das, P.; Mulherkar, R.; Singh, R.; Dutt, A. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br. J. Cancer* **2015**, *112*, 1958–1965. [CrossRef] [PubMed]
43. Zhang, R.; Shen, C.; Zhao, L.; Wang, J.; McCrae, M.; Chen, X.; Lu, F. Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis. *Int. J. Cancer* **2016**, *138*, 1163–1174. [CrossRef] [PubMed]
44. Ibragimova, M.; Tsyganov, M.; Shpileva, O.; Churuksaeva, O.; Bychkov, V.; Kolomiets, L.; Litviakov, N. HPV status and its genomic integration affect survival of patients with cervical cancer. *Neoplasia* **2018**, *65*, 441–448. [CrossRef] [PubMed]
45. Shen, C.; Liu, Y.; Shi, S.; Zhang, R.; Zhang, T.; Xu, Q.; Zhu, P.; Chen, X.; Lu, F. Long-distance interaction of the integrated HPV fragment with MYC gene and 8q24.22 region upregulating the allele-specific MYC expression in HeLa cells. *Int. J. Cancer* **2017**, *141*, 540–548. [CrossRef]
46. Koneva, L.A.; Zhang, Y.; Virani, S.; Hall, P.B.; McHugh, J.B.; Chepeha, D.B.; Wolf, G.T.; Carey, T.E.; Rozek, L.S.; Sartor, M.A. HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers. *Mol. Cancer Res.* **2018**, *16*, 90–102. [CrossRef]
47. Han, L.; Maimaitiming, T.; Husaiyin, S.; Wang, L.; Wusainahong, K.; Ma, C.; Niyazi, M. Comparative study of HPV16 integration in cervical lesions between ethnicities with high and low rates of infection with high-risk HPV and the correlation between integration rate and cervical neoplasia. *Exp. Ther. Med.* **2015**, *10*, 2169–2174. [CrossRef] [PubMed]



48. Liu, L.; Ying, C.; Zhao, Z.; Sui, L.; Zhang, X.; Qian, C.; Wang, Q.; Chen, L.; Guo, Q.; Wu, J. Identification of reliable biomarkers of human papillomavirus 16 methylation in cervical lesions based on integration status using high-resolution melting analysis. *Clin. Epigenetics* **2018**, *10*, 10. [CrossRef] [PubMed]
49. Jiang, Y.; Zhu, C.; He, D.; Gao, Q.; Tian, X.; Ma, X.; Wu, J.; Das, B.C.; Severinov, K.; Hitzeroth, I.I.; et al. Cytological Immunostaining of HMGA2, LRP1B, and TP63 as Potential Biomarkers for Triaging Human Papillomavirus-Positive Women. *Transl. Oncol.* **2019**, *12*, 959–967. [CrossRef]
50. Tuna, M.; Amos, C.I. Next generation sequencing and its applications in HPV-Associated cancers. *Oncotarget* **2017**, *8*, 8877–8889. [CrossRef]
51. Chae, J.; Park, W.S.; Kim, M.J.; Jang, S.S.; Hong, D.; Ryu, J.; Ryu, C.H.; Kim, J.H.; Choi, M.K.; Cho, K.H.; et al. Genomic characterization of clonal evolution during oropharyngeal carcinogenesis driven by human papillomavirus 16. *BMB Rep.* **2018**, *51*, 584–589. [CrossRef]
52. Yang, W.; Liu, Y.; Dong, R.; Liu, J.; Lang, J.; Yang, J.; Wang, W.; Li, J.; Meng, B.; Tian, G. Accurate Detection of HPV Integration Sites in Cervical Cancer Samples Using the Nanopore MinION Sequencer Without Error Correction. *Front. Genet.* **2020**, *11*, 660. [CrossRef] [PubMed]
53. Arroyo-Muhr, L.S.; Lagheden, C.; Hultin, E.; Eklund, C.; Adami, H.O.; Dillner, J.; Sundstrom, K. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: Prospective population-based study. *Br. J. Cancer* **2018**, *119*, 1163–1168. [CrossRef]
54. Blanco, L.; Bernad, A.; Lazaro, J.M.; Martin, G.; Garmendia, C.; Salas, M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **1989**, *264*, 8935–8940. [CrossRef]
55. Binga, E.K.; Lasken, R.S.; Neufeld, J.D. Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. *ISME J.* **2008**, *2*, 233–241. [CrossRef]
56. Polson, S.W.; Wilhelm, S.W.; Wommack, K.E. Unraveling the viral tapestry (from inside the capsid out). *ISME J.* **2011**, *5*, 165–168. [CrossRef] [PubMed]
57. Li, T.; Unger, E.R.; Rajeevan, M.S. Universal human papillomavirus typing by whole genome sequencing following target enrichment: Evaluation of assay reproducibility and limit of detection. *BMC Genom.* **2019**, *20*, 231. [CrossRef]
58. Tjalma, W. HPV negative cervical cancers and primary HPV screening. *Facts Views Vis. Obgyn* **2018**, *10*, 107–113.
59. Walboomers, J.M.; Jacobs, M.V.; Manos, M.M.; Bosch, F.X.; Kummer, J.A.; Shah, K.V.; Snijders, P.J.; Peto, J.; Meijer, C.J.; Munoz, N. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **1999**, *189*, 12–19. [CrossRef]
60. Lei, J.; Ploner, A.; Lagheden, C.; Eklund, C.; Nordqvist Kleppe, S.; Andrae, B.; Elfstrom, K.M.; Dillner, J.; Sparen, P.; Sundstrom, K. High-risk human papillomavirus status and prognosis in invasive cervical cancer: A nationwide cohort study. *PLoS Med.* **2018**, *15*, e1002666. [CrossRef] [PubMed]
61. De Sanjose, S.; Quint, W.G.; Alemany, L.; Geraets, D.T.; Klaustermeier, J.E.; Lloveras, B.; Tous, S.; Felix, A.; Bravo, L.E.; Shin, H.R.; et al. Human papillomavirus genotype attribution in invasive cervical cancer: A retrospective cross-sectional worldwide study. *Lancet Oncol.* **2010**, *11*, 1048–1056. [CrossRef]
62. Arroyo Muhr, L.S.; Lagheden, C.; Eklund, C.; Lei, J.; Nordqvist-Kleppe, S.; Sparen, P.; Sundstrom, K.; Dillner, J. Sequencing detects human papillomavirus in some apparently HPV-negative invasive cervical cancers. *J. Gen. Virol.* **2020**, *101*, 265–270. [CrossRef]
63. Gillison, M.L.; D'Souza, G.; Westra, W.; Sugar, E.; Xiao, W.; Begum, S.; Viscidi, R. Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. *J. Natl. Cancer Inst.* **2008**, *100*, 407–420. [CrossRef]
64. Wakeham, K.; Kavanagh, K.; Cuschieri, K.; Millan, D.; Pollock, K.G.; Bell, S.; Burton, K.; Reed, N.S.; Graham, S.V. HPV status and favourable outcome in vulvar squamous cancer. *Int. J. Cancer* **2017**, *140*, 1134–1146. [CrossRef] [PubMed]
65. O'Sullivan, B.; Huang, S.H.; Su, J.; Garden, A.S.; Sturgis, E.M.; Dahlstrom, K.; Lee, N.; Riaz, N.; Pei, X.; Koyfman, S.A.; et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): A multicentre cohort study. *Lancet Oncol.* **2016**, *17*, 440–451. [CrossRef]
66. World Health Organization. Female Genital Tumors. *WHO Classification of Tumors*, 5th ed. Volume 5. Available online: <https://publications.iarc.fr/592> (accessed on 4 April 2021).
67. Lubock, N.B.; Zhang, D.; Sidore, A.M.; Church, G.M.; Kosuri, S. A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res.* **2017**, *45*, 9206–9217. [CrossRef]
68. Mitchell, K.; Brito, J.J.; Mandric, I.; Wu, Q.; Knyazev, S.; Chang, S.; Martin, L.S.; Karlsberg, A.; Gerasimov, E.; Littman, R.; et al. Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* **2020**, *21*, 71. [CrossRef] [PubMed]
69. Lindgreen, S. AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **2012**, *5*, 337. [CrossRef]
70. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [CrossRef] [PubMed]
71. Broad Institute. Genome Analysis Toolkit. Available online: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores> (accessed on 26 May 2021).
72. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.20133997.
73. Li, R.; Yu, C.; Li, Y.; Lam, T.W.; Yiu, S.M.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967. [CrossRef]

74. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
75. Novoalign. Available online: <http://novocraft.com/> (accessed on 26 May 2021).
76. Sedlazeck, F.J.; Rescheneder, P.; Von Haeseler, A. NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **2013**, *29*, 2790–2791. [CrossRef]
77. Arroyo Muhr, L.S.; Eklund, C.; Dillner, J. Misclassifications in human papillomavirus databases. *Virology* **2021**, *558*, 57–66. [CrossRef] [PubMed]
78. Ekstrom, J.; Muhr, L.S.; Bzhalava, D.; Soderlund-Strand, A.; Hultin, E.; Nordin, P.; Stenquist, B.; Paoli, J.; Forslund, O.; Dillner, J. Diversity of human papillomaviruses in skin lesions. *Virology* **2013**, *447*, 300–311. [CrossRef]
79. de Villiers, E.M.; Fauquet, C.; Broker, T.R.; Bernard, H.U.; zur Hausen, H. Classification of papillomaviruses. *Virology* **2004**, *324*, 17–27. [CrossRef] [PubMed]
80. Bernard, H.U.; Burk, R.D.; Chen, Z.; Van Doorslaer, K.; Zur Hausen, H.; De Villiers, E.M. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **2010**, *401*, 70–79. [CrossRef]
81. De Villiers, E.M. Cross-roads in the classification of papillomaviruses. *Virology* **2013**, *445*, 2–10. [CrossRef] [PubMed]
82. Burk, R.D.; Harari, A.; Chen, Z. Human papillomavirus genome variants. *Virology* **2013**, *445*, 232–243. [CrossRef] [PubMed]
83. Smith, B.; Chen, Z.; Reimers, L.; Van Doorslaer, K.; Schiffman, M.; Desalle, R.; Herrero, R.; Yu, K.; Wacholder, S.; Wang, T.; et al. Sequence imputation of HPV16 genomes for genetic association studies. *PLoS ONE* **2011**, *6*, e21375. [CrossRef]
84. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
85. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
86. Kumar, S.; Stecher, G.; Li, M.; Nnyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
87. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
88. International Journal of Cancer. Submission Guidelines. Available online: [https://onlinelibrary.wiley.com/pb-assets/assets/10970215/IJC\\_Sequencing\\_Coverage\\_and\\_Quality\\_Statistics\\_Guidelines-1607431877843.pdf](https://onlinelibrary.wiley.com/pb-assets/assets/10970215/IJC_Sequencing_Coverage_and_Quality_Statistics_Guidelines-1607431877843.pdf) (accessed on 26 May 2021).
89. Poljak, M.; Cuschieri, K.; Waheed, D.E.; Baay, M.; Vorsters, A. Impact of the COVID-19 pandemic on human papillomavirus-based testing services to support cervical cancer screening. *Acta Dermatovenerol. Alp. Pannonica Adriat.* **2021**, *30*, 21–26.
90. Hilke, F.J.; Muiyas, F.; Admard, J.; Kootz, B.; Nann, D.; Welz, S.; Riess, O.; Zips, D.; Ossowski, S.; Schroeder, C.; et al. Dynamics of cell-free tumour DNA correlate with treatment response of head and neck cancer patients receiving radiochemotherapy. *Radiother. Oncol.* **2020**, *151*, 182–189. [CrossRef]
91. Wagner, S.; Roberson, D.; Boland, J.; Yeager, M.; Cullen, M.; Mirabello, L.; Dunn, S.T.; Walker, J.; Zuna, R.; Schiffman, M.; et al. Development of the TypeSeq Assay for Detection of 51 Human Papillomavirus Genotypes by Next-Generation Sequencing. *J. Clin. Microbiol.* **2019**, *57*. [CrossRef] [PubMed]
92. Chan, W.S.; Chan, T.L.; Au, C.H.; Leung, C.P.; To, M.Y.; Ng, M.K.; Leung, S.M.; Chan, M.K.M.; Ma, E.S.K.; Tang, B.S.F. An economical Nanopore sequencing assay for human papillomavirus (HPV) genotyping. *Diagn. Pathol.* **2020**, *15*, 45. [CrossRef]
93. Lahens, N.F.; Ricciotti, E.; Smirnova, O.; Toorens, E.; Kim, E.J.; Baruzzo, G.; Hayer, K.E.; Ganguly, T.; Schug, J.; Grant, G.R. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genom.* **2017**, *18*, 602. [CrossRef]
94. Marine, R.L.; Magana, L.C.; Castro, C.J.; Zhao, K.; Montmayeur, A.M.; Schmidt, A.; Diez-Valcarce, M.; Ng, T.F.F.; Vinje, J.; Burns, C.C.; et al. Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. *J. Virol. Methods* **2020**, *280*, 113865. [CrossRef]
95. Hornung, B.V.H.; Zwitter, R.D.; Kuijper, E.J. Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* **2019**, *95*. [CrossRef] [PubMed]
96. Zur Hausen, H. Papillomaviruses and cancer: From basic studies to clinical application. *Nat. Rev. Cancer* **2002**, *2*, 342–350. [CrossRef] [PubMed]
97. Bzhalava, Z.; Arroyo Muhr, L.S.; Dillner, J. Transcription of human papillomavirus oncogenes in head and neck squamous cell carcinomas. *Vaccine* **2020**, *38*, 4066–4070. [CrossRef] [PubMed]
98. Braakhuis, B.J.; Snijders, P.J.; Keune, W.J.; Meijer, C.J.; Ruijter-Schippers, H.J.; Leemans, C.R.; Brakenhoff, R.H. Genetic patterns in head and neck cancers that contain or lack transcriptionally active human papillomavirus. *J. Natl. Cancer Inst.* **2004**, *96*, 998–1006. [CrossRef]
99. Leemans, C.R.; Braakhuis, B.J.; Brakenhoff, R.H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **2011**, *11*, 9–22. [CrossRef]
100. World Health Organization. Available online: <http://www.who.int/biologicals/vaccines/hpv/en/index.htm> (accessed on 26 May 2021).
101. World Health Organization. *Human Papillomavirus Laboratory Manual*, 1st ed.; Geneva, World Health Organization: Geneva, Switzerland, 2009. Available online: <https://apps.who.int/iris/handle/10665/70505> (accessed on 4 April 2021).

102. Eklund, C.; Forslund, O.; Wallin, K.L.; Dillner, J. Continuing global improvement in human papillomavirus DNA genotyping services: The 2013 and 2014 HPV LabNet international proficiency studies. *J. Clin. Virol.* **2018**, *101*, 74–85. [[CrossRef](#)] [[PubMed](#)]
103. Cuschieri, K.; Schuurman, R.; Coughlan, S. Ensuring quality in cervical screening programmes based on molecular human papillomavirus testing. *Cytopathology* **2019**, *30*, 273–280. [[CrossRef](#)] [[PubMed](#)]
104. Fagan, E.J.; Moore, C.; Jenkins, C.; Rossouw, A.; Cubie, H.A.; James, V.L. External quality assessment for molecular detection of human papillomaviruses. *J. Clin. Virol.* **2010**, *48*, 251–254. [[CrossRef](#)] [[PubMed](#)]

## Publication 2

Received: 31 December 2021 | Revised: 22 March 2022 | Accepted: 6 April 2022

DOI: 10.1002/cam4.4771

RESEARCH ARTICLE

Cancer Medicine  WILEY

# HPV status and HPV16 viral load in anal cancer and its association with clinical outcome

Daniel Guerendiain<sup>1,2</sup>  | Raluca Grigorescu<sup>3</sup> | Anna Kirk<sup>4</sup> | Andrew Stevenson<sup>4</sup> | Matthew T. G. Holden<sup>2</sup> | Jiafeng Pan<sup>5</sup>  | Kim Kavanagh<sup>5</sup> | Sheila V. Graham<sup>4</sup> | Kate Cuschieri<sup>1</sup>

<sup>1</sup>Scottish HPV Reference Laboratory, NHS Lothian, Edinburgh, UK

<sup>2</sup>School of Medicine, University of St Andrews, St Andrews, UK

<sup>3</sup>Pathology, NHS Lothian, Edinburgh, UK

<sup>4</sup>Centre for Virus Research, Institute of Infection Immunity and Inflammation, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

<sup>5</sup>Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK

### Correspondence

Daniel Guerendiain, Scottish HPV Reference Laboratory, Royal Infirmary of Edinburgh, 51 Little France Crescent, EH16 4SA, UK and School of Medicine, University of St Andrews, St Andrews, UK.  
Email: [daniel.guerendiain@nhslothian.scot.nhs.uk](mailto:daniel.guerendiain@nhslothian.scot.nhs.uk)

### Abstract

**Background:** The incidence of anal cancer is increasing globally. Evidence-based improvement in early detection and management of this morbid cancer is thus required. In other cancers associated with Human Papillomavirus (HPV), viral status and dynamics, including viral load (VL) has been shown to influence clinical outcome. Our aim was to determine the influence of HPV status and HPV16 VL on the clinical outcomes of anal cancer patients.

**Methods:** A total of 185 anal cancer lesions were genotyped for HPV. Of the HPV16 positive component, VL was determined using a digital droplet PCR assay. The association of qualitative HPV status and VL (low (<12.3), medium (12.3–57) and high (>57 copies/cell)) on overall survival and hazard of death was assessed.

**Results:** Of the 185 cases, 164 (88.6%) samples were HPV positive. HPV16 was detected in 154/185 samples (83.2%). HPV positive status was associated with improved overall survival in the univariate analysis [hazard ratio (HR) of 0.44, 0.23–0.82,  $p = 0.01$ ]. When adjusted by age, sex, stage and response to treatment, the association of positive HPV status with improved survival remained (HR 0.24 [0.11–0.55]  $p < 0.001$ ). High VL was associated with improved overall survival in the univariate analysis with a HR of 0.28 (0.11–0.71,  $p = 0.007$ ). When adjusted only by age and sex, high VL was associated with better overall survival (HR 0.27, 0.11–0.68  $p = 0.006$ ).

**Conclusions:** HPV status appears to be independently associated with improved outcomes in anal cancer patients. Moreover, HPV viral load quantification may be informative for further risk stratification and warrants further investigation.

### KEYWORDS

anal cancer, clinical outcome, viral load

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

*Cancer Medicine*. 2022;11:4193–4203.

[wileyonlinelibrary.com/journal/cam4](https://wileyonlinelibrary.com/journal/cam4) | 4193



## 1 | INTRODUCTION

Anal cancer is one of the six cancers shown to have a human papillomavirus (HPV) aetiology with approximately 90% of anal cancers being HPV-driven.<sup>1,2</sup> Like oropharyngeal cancer, there appears to be a clear dominant HPV type driving the lesion. In the meta-analysis study carried out by De Sanjosé et al. 2019, HPV16 was present in 80.7% of anal cancers.<sup>3</sup>

As with other HPV-driven cancers, anal cancer incidence is increasing worldwide, including in the USA and Europe.<sup>4,7</sup> Scottish European age-standardised rate (EASR) (per 100,000 person-years at risk) data align with these increases, rising from 1.5 in 1995 to 2.6 in 2017. For men, incidence increased from 1.6 in 1995 to 2.1 in 2017 while in females, incidence increased from 1.2 in 1995 to 3 in 2017.<sup>8</sup> The rest of the UK has also experienced an increase for both sexes from 1.5 in 1993 to 1.7 in 2017 in males and 1.3 to 3.0 in females over the same period.<sup>6</sup>

Several factors have been linked to an increased risk of anal cancer including age (higher proportion of cases occur in people after 50 years of age), number of sexual partners, history of receptive anal sexual intercourse, smoking and immune capacity.<sup>9,10</sup> Women who have or had high-grade or worse cervical lesions (CIN2/3+) and/or vulvar high-grade lesions also have a higher risk of anal lesions and cancer.<sup>11-13</sup> Moreover, incidence of anal cancer is significantly higher in people living with HIV (PLWH).<sup>14-16</sup>

The morbidity associated with anal lesion and cancer treatments can be very high. Treatments include chemoradiotherapy (CRT) or resection of the affected tissue, depending on the area affected (anal margin cancers are treated in a slightly different way from anal canal tumours).<sup>17,18</sup> Treatment can have a significant deleterious impact in the patient's quality of life including issues with sexual function and faecal continence.<sup>18-20</sup>

Various studies have assessed HPV viral load (VL) in the different HPV-driven cancers and its association with overall survival and prognosis and/or as a biomarker for lesion progression. In cervical cancers, low HPV viral load may implicate a worse prognosis (median value of 385.8 RLU/CO Kim et al. 2009<sup>21</sup> 132.5 RLU/CO in Deng et al.<sup>22</sup>). The number of copies of the HPV genome has also been examined in oropharyngeal cancer (OPC) cases driven by HPV and a number of studies have observed that high viral load is associated with a better prognosis versus low viral load.<sup>23-27</sup> Additionally, disease recurrence has been shown to be significantly lower in those with high HPV load.<sup>23</sup> For these publications median value ranged between 30.9, 132.5, 190 and 820 copies/cell.

Complimentary studies have investigated the implications of HPV viral load in and around the anal canal,<sup>28-32</sup>

however, the majority of the available publications have looked into HPV load in people living with HIV. Other studies have assessed VL in anal cancer<sup>31,32</sup> and found a median of 7.40 and 134 copies/cell. However, to our knowledge only one investigation has explored the association between HPV load, local control of cancer and overall survival.<sup>31</sup> Authors found that patients with HPV16 DNA VL below the median viral load with low p16 expression showed significantly worse local control and overall survival (OS) than those with a VL above median.<sup>32</sup>

The reason behind why cases with HPV high viral load have better prognosis or OS is not completely understood, although it has been posited that it could be associated with the HPV episome status and integration as episome is associated with higher copies of HPV.<sup>25,33-35</sup>

As is the case in most settings, there is no population-based anal screening programme in place in Scotland (where the present work was performed) and generally, anal lesions are detected and managed, clinically. Moreover, as HPV testing is not currently included in the diagnostic work up of anal disease, there is no routinely collected information on HPV-associated epidemiology in anal disease in Scotland. Proactive efforts are therefore needed to understand the nature and implications of HPV in anal cancer to inform prevention and management strategies of the future.

Our aim was to determine the HPV-attributable fraction of anal cancers, that is, the fraction associated with HPV in Scotland, and to use this data to ascertain the influence of qualitative HPV status on clinical outcomes. Then, using this well annotated data set we also aimed to determine whether VL (in the HPV16 positive component) exerted an influence on clinical outcomes.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample collation and timeframe

Squamous cell carcinoma of anus or anal canal samples (ICD-11 coding 2C00.3) preserved in formalin fixed paraffin embedded (FFPE) blocks diagnosed between 2009 and 2018 were collated. These samples were originally obtained as part of standard of care for the management of patients with anal disease.<sup>36</sup> All biopsies were obtained from the South-east of Scotland representing 3 of 14 territorial health boards in Scotland; NHS Lothian, NHS Borders and NHS Fife. These health boards serve a population of 1,396,640 (data from 2019).<sup>37</sup> Favourable ethical opinion to conduct the research was provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482. This was further

supported by approval for use of samples through the National Research for Scotland Bioresource (20/ES/0061), application reference SR1283.

Clinico-demographic information was obtained on age, sex, stage of cancer (using the American Joint Committee on Cancer [AJCC] TNM system),<sup>38</sup> response to treatment, date of diagnosis and vital (dead/alive) status. Age and stage of cancer were considered at the time of diagnosis. Information was obtained in January 2020 and indexed with a study number. Vital status information and date of death data were censored in July 2020. Response to treatment was aggregated in two different groups to simplify the analysis: response to treatment (including remission) and no response to treatment (including progression and recurrence). Cohort follow-up started at date of diagnosis and continued until death or time of censoring.

Cases categorised according to the various clinical and demographic variables are summarised in Table 1. Age was stratified in four different groups: <50, 50–59, 60–69 and ≥70. Response to treatment was organised in three groups: yes, no or unknown following the ESMO guidelines for anal cancer.<sup>39</sup> Cancer stage was aggregated in five groups: I, II, III, IV and unknown following AJCC system effective January 2018.<sup>38</sup>

## 2.2 | Overarching approach to HPV annotation: Qualitative analysis and quantitative analysis

A total of 185 anal cancer samples were annotated for HPV type-specific prevalence, initially, using a PCR-based assay: the Anyplex II 28 assay (Seegene, Korea) centrally at the Scottish HPV Reference Laboratory, Edinburgh, UK. One 10 µm section per sample was obtained and incubated in Seegene Universal Lysis Buffer (LB) at 65°C overnight. DNA extraction was performed using the MicroLab Nimbus IVD (Hamilton) with the StarMag Universal cartridge Kit (Seegene). Mastermix was prepared with the Nimbus and PCR on the CFX Real-time PCR instrument (Biorad).

This initial result allowed determination of type-specific prevalence and examination of the association between qualitative HPV status (any HPV vs. no HPV) and survival, which we subsequently term the 'qualitative analysis'. Given the dominance of HPV16 in the cohort, viral load analysis was restricted to samples that tested HPV16 positive, either as a mono infection or within a mixed infection. Once viral load was obtained (quantitative analysis), it was classified in three different groups: low, medium and high and linked to overall survival. A diagram illustrating the process followed can be found in Appendix A.

TABLE 1 Demographic & clinical characteristics of the anal cancer patients collected between 2009 and 2018 in the South-East of Scotland. Stratification by HPV status and HPV16 Viral load

Variable	Level	HPV status & survival samples cohort N (%)			HPV16 Viral load & survival samples cohort N (%)			
		n = 185	HPV + ve (n = 164)	HPV - ve (n = 21)	n = 145	Low (n = 47)	Medium (n = 50)	High (n = 48)
Sex	Female	120 (64.9%)	109 (66.5%)	11 (52.4%)	101 (69.7%)	29 (61.7%)	38 (76.0%)	34 (70.8%)
	Male	65 (35.1%)	55 (33.5%)	10 (47.6%)	44 (30.3%)	18 (38.3%)	12 (24.0%)	14 (29.2%)
Age	<50	28 (15.1%)	25 (15.2%)	3 (14.3%)	22 (15.2%)	7 (14.9%)	6 (12.0%)	9 (18.8%)
	50–59	48 (25.9%)	48 (29.3%)	0 (0%)	44 (30.3%)	18 (38.3%)	14 (28.0%)	12 (25.0%)
	60–69	56 (30.3%)	51 (31.1%)	5 (23.8%)	45 (31.0%)	17 (36.2%)	12 (24.0%)	16 (33.3%)
	70 and over	53 (28.6%)	40 (24.4%)	13 (61.9%)	34 (23.4%)	5 (10.6%)	18 (36.0%)	11 (22.9%)
Stage	I	27 (14.6%)	23 (14.0%)	4 (19.0%)	22 (15.2%)	5 (10.6%)	8 (16.0%)	9 (18.8%)
	II	68 (36.8%)	62 (37.8%)	6 (28.6%)	54 (37.2%)	15 (31.9%)	17 (34.0%)	22 (45.8%)
	III	53 (28.6%)	48 (29.3%)	5 (23.8%)	40 (27.6%)	15 (31.9%)	16 (32.0%)	11 (22.9%)
	IV	35 (18.9%)	23 (14.0%)	6 (28.6%)	25 (17.2%)	11 (23.4%)	9 (18.0%)	5 (10.4%)
	Unknown	2 (1.1%)	2 (1.2%)	0 (0.0%)	2 (1.4%)	1 (2.1%)	0 (0.0%)	1 (2.1%)
Response to Treatment	Yes	138 (74.6%)	125 (76.2%)	13 (61.9%)	111 (76.6%)	35 (74.5%)	37 (74.0%)	39 (81.2%)
	No	34 (18.4%)	28 (17.1%)	6 (28.6%)	23 (15.9%)	8 (17.0%)	9 (18.0%)	6 (12.5%)
	Unknown	13 (7.0%)	11 (6.7%)	2 (9.5%)	11 (7.6%)	4 (8.5%)	4 (8.0%)	3 (6.3%)
Vital status	Alive	124 (67.0%)	115 (70.1%)	9 (42.9%)	104 (71.7%)	29 (61.7%)	33 (66.0%)	42 (87.5%)
	Deceased	61 (33.0%)	49 (29.9%)	12 (57.1%)	41 (28.3%)	18 (38.3%)	17 (34.0%)	6 (12.5%)

Note: 'N's corresponds to the total number of samples for every category described above. Percentage (%) was calculated from the total number of valid samples (n = 185 & 145).



### 2.3 | Measurement of viral load using ddPCR

Absolute quantification of viral load was performed on HPV16+ve cancer samples (145 mono and nine mixed infections) using a droplet digital assay (ddPCR). This element of our analysis is referred to as the quantitative analysis group. Nucleic acid was extracted from HPV16+ve samples (new section, 10 µm) using the QIAamp DNA Mini Kit (Qiagen) and sample concentration measurement was performed with the Qubit dsDNA High sensitivity kit (Thermo Fisher Scientific).

ddPCR was performed as described in Stevenson et al. 2020<sup>23</sup> at the Centre for Virus Research, University of Glasgow. The RPP30 endogenous control probe primer set of 0.7 µl, HPV16 L1-specific primers and probes at 300 nM (final concentration) respectively, 10–100 ng of template DNA and 1 µl of restriction digest mix (consisting of 4 U of both *EcoRI* and *HindIII* in 1x NEB Cutsmart buffer [NEB, UK]) were used for the mix. Reactions were mixed with Droplet Generation Oil on DG8 cartridges in the QX200 droplet generator (Bio-Rad) to generate droplets. Thermal cycling conditions were: 95°C for 10 min followed by 40 × 30s at 94°C and 60°C for 1 min prior to final extension at 98°C for 10 min.

Post-amplification, droplets were analysed on a QX200 Droplet Reader (Bio-Rad), and output data files were analysed using QuantaSoft analysis software v1.7.4 (Bio-Rad). The viral load for each sample was calculated relative to the endogenous RRP30 cellular gene internal control, with two copies present per cell. Any initially invalid results were repeated using a new FFPE section and fresh DNA extraction. After retesting, consistent invalids were not included in the analysis ( $n = 9$ ).

### 2.4 | Definition of viral load levels

The individual HPV16 viral loads were ranked from smallest to largest and separated using tertiles. The VL threshold(s) for L1 low viral load was <12.3, medium between 12.3 and 57 and high viral load above 57 copies/cell.

### 2.5 | Association of HPV status and viral load with demographics and survival outcomes

Overall survival by qualitative status of HPV (HPV positive vs. negative; HPV16 positive vs. HPV negative) and HPV VL (low, medium and high) was analysed using the Kaplan–Meier method. The univariate and

multivariate hazard ratios of HPV status (negative vs. positive) and virus load (low vs. medium and high) for all cause death were derived using Cox proportional hazard model. Two multivariate models were derived—age (<50, 50–59, 60–69, 70+) and sex adjusted model as described in Stevenson et al.<sup>23</sup> and a fully adjusted model, where age, sex, stage (I, II, III, IV) and response to treatment (no, yes) were adjusted for. All the statistical analysis were performed using R-studio (version 1.2.1335).<sup>40–42</sup>

For the qualitative detection HPV, the types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68 were considered as high risk. HPV types considered as low-risk HPV types include: 6, 11, 40, 42, 43, 44, 54 and 61.

## 3 | RESULTS

### 3.1 | Clinical/demographic characterisation of cohort

Overall HPV status stratified by demographic and clinical characteristics are presented in Table 1. The cohort contained 64.9% samples from females, and 35.1% males. The majority of cases were diagnosed in individuals aged 60–69 (30.3%) and the majority of cases were stage II and III (36.7% and 28.6%). Additionally, 74.6% of cases responded to treatment and 67.0% were alive at date of censoring.

Of the female cases 90.8% were HPV positive; of the male cases, 84.6% were HPV positive. The majority of HPV positive cases were diagnosed in the 60–69 (31.1%) age range and were stage II (37.8%). Overall, 76.2% responded to treatment and 70.1% were alive at censoring.

With respect to the cases assessed for the quantitative analysis, 69.7% of samples were from females, 30.3% from males. The majority were diagnosed aged 60–69 years (31.0%), at stage II (37.2%). A total of 76.6% responded to treatment and 71.7% were alive at time of censoring. A higher proportion of those with high VL responded to treatment compared to those with low/medium VL (81.2% for the high VL group vs. 74.5% and 74.0% for the low and medium VL group). A higher proportion of those with high VL were alive at the time of censoring (87.5% for the high VL group vs. 61.7% and 66.0% for the low and medium VL group. (Table 1).

### 3.2 | Qualitative analysis of HPV in anal cancer samples

A total of 185 anal cancer samples were genotyped for HPV. Of the 185 cases, 164 (88.6%) samples were

positive for at least one HPV type. High-risk (hr) types were detected in 87.03% of the samples. Monoinfection of HPV16 was present in 145 (78.4%) samples. Seven samples (3.8%) had a combination of HPV16 and other hr-HPV type(s). Two samples (1.1%) were co infected with HPV16 and a low-risk type. HPV18 was the second most dominant type, present in three samples (1.6%). HPV 33 and 68 were detected in two samples (1.1%), HPV 35, HPV 51 and HPV 52 in 1 (0.5%) while HPV 39 was detected in three (1.6%). The presence of low-risk types without any other hr-HPV was detected in three samples (1.6%).

### 3.3 | Does HPV positivity have an impact on survival in patients with anal cancer?

Of the 185 cases included in the qualitative analysis, 61 (33.0%) patients died during follow-up.

Kaplan–Meier curves were produced and stratified by HPV status (positive and negative) (Figure 1A), and HPV16 status (Figure 1B). HPV positivity and HPV16 positive status were associated with better survival (log-rank test *p* value 0.0077 and 0.006, respectively).

HPV +ve status was associated with improved overall survival in the univariate analysis with a hazard ratio

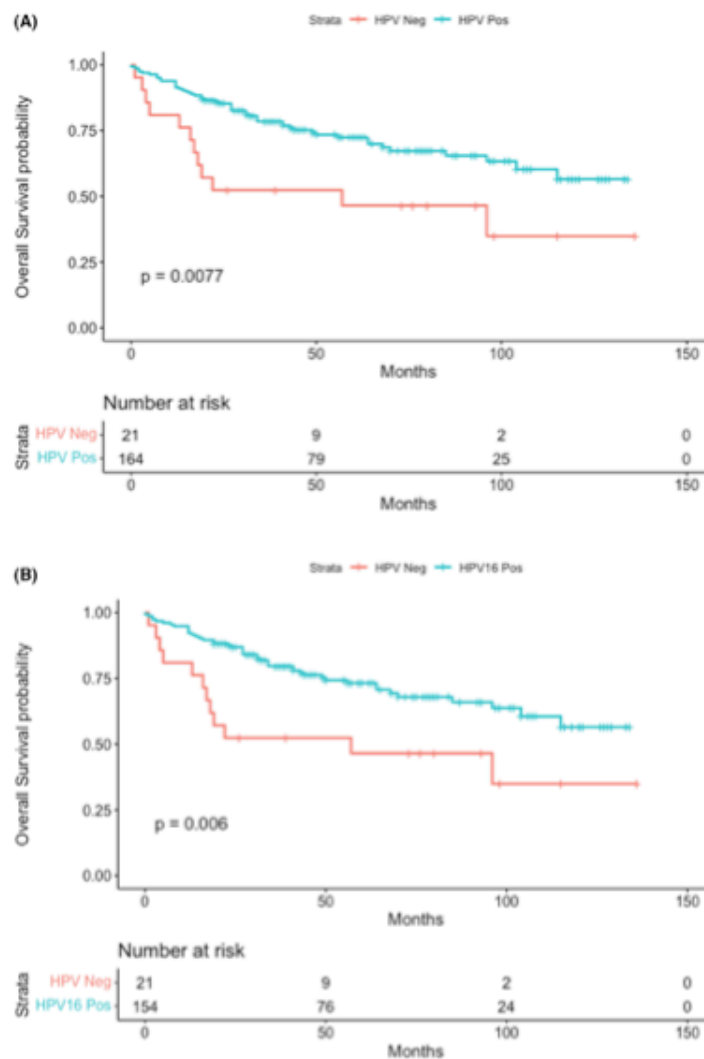


FIGURE 1 Overall Survival probability for 'any' HPV positive versus HPV negative cases (A) and for HPV16 positive and non-HPV16 positive cases (B) using Kaplan–Meier estimator. Survival time is expressed in months from diagnosis date. Data censored in July 2020

(HR) of 0.44 (0.23–0.82,  $p = 0.01$ ) (Table 2). In the univariate Cox model, variables associated with worse overall survival were Stage III; HR 5.0 (1.1–22),  $p = 0.003$  and Stage IV HR 25.6 (6.0–109),  $p < 0.001$  versus stage I and response to treatment 0.12 (0.07–0.33)  $p < 0.001$  versus non response to treatment. After adjusting for age, gender, stage and response to treatment, HPV status continued to influence the overall survival, HR 0.24 (0.11–0.55)  $p < 0.001$ . When adjusting for age and gender alone, HR for HPV positive status was 0.41(0.21–0.82)  $p = 0.011$ .

### 3.4 | Viral load range in the anal cancer samples

A total of 145/154 HPV16 positive samples (94.1%) were associated with valid reads in the ddPCR for HPV16 L1 sequences. Nine samples were excluded from the analysis because they generated less than 10,000 droplets (the threshold for validity) even after repeat testing.

Viral loads ranged from 0.021 to 710 copies of the HPV L1 gene per cell with a mean of 60.57 L1 copies. Those who were deceased at time of analysis (41/145, 28.28%) had a median L1 VL of 33.11 (IQR 3.6–43.5); while those still alive (104/145, 71.72%) had a median L1 VL of 74.57 (IQR 8.2–104.5) (Table 3). A total of 47 samples (32.4%) had a low VL, 50 a medium VL (34.5%) and 48 (33.1%) a high VL. Mean VL was 3.8, 35.6 and 148.9 for low, medium and high VL groups, respectively. Viral load stratified by vital status, irrespective of underlying demographics is described in Table 3.

### 3.5 | Viral load and impact on clinical outcomes

Of those alive at the time of data censoring, 27.9% cancer samples were associated with a low VL, 31.7% a medium VL and 40.4% a high VL. Comparatively, in those who died low VL was present in 43.9%, medium VL in 41.5% whereas 14.6% had a high VL (Table 3).

For the Kaplan–Meier estimator, overall survival was calculated by classifying viral load in three groups, low, medium and high. Overall survival in those with medium and high viral load was higher than in those with low VL ( $p = 0.026$ ), Figure 2.

Table 4 shows overall survival stratified by the clinical and demographic variables described in Table 1 with viral load categorised into the three tertiles with low VL as the reference. High viral load was associated with improved overall survival in the univariate analysis with a hazard ratio (HR) of 0.28 (0.11–0.71,  $p = 0.007$ ) compared to low viral load. Variables associated with worse overall survival in the univariate model were Stage IV vs. stage I with HR of 25.2 (5.65–113),  $p < 0.001$  and response to treatment [HR of 0.13 (0.064–0.27)  $p < 0.001$ ] versus non response to treatment.

After adjustment for age, gender, stage and response to treatment, viral load did not significantly influence the overall survival; medium VL HR 1.04 (0.45–2.40)  $p = 0.924$ , high VL 0.39 (0.12–1.24)  $p = 0.111$  compared to low VL. In the age/gender adjusted Cox model, high viral load was still associated with improved overall survival compared to low VL (0.27, 0.11–0.68)  $p = 0.006$ .

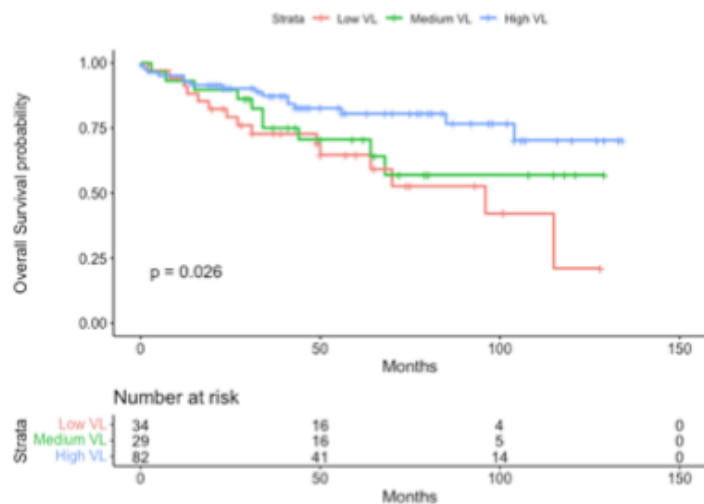
TABLE 2 Univariate and multivariate hazard ratio of HPV status derived using Cox regression ( $N = 185$ )

Variable	Level	Unadjusted HR (95% Cis)	$p$ value	Adjusted HR (95% Cis)	$p$ value	Adjusted HR (95% Cis)	$p$ value
HPV	HPV Neg	1		1		1	
	HPV Pos	0.44 (0.23–0.82)	0.01	0.24 (0.11–0.55)	<0.001	0.41 (0.21–0.82)	0.011
Sex	Male	1		1		1	
	Female	0.85 (0.51–1.4)	0.549	0.98 (0.52–1.87)	0.955	0.90 (0.53–1.53)	0.704
Age	<50	1		1		1	
	50–59	1.6 (0.68–3.7)	0.288	1.09 (0.43–2.79)	0.852	1.84 (0.77–4.37)	0.167
	60–69	1.2 (0.53–2.9)	0.635	2.40 (0.97–5.98)	0.059	1.31 (0.56–3.06)	0.532
	70 and over	1.6 (0.70–3.7)	0.257	1.88 (0.69–5.11)	0.217	1.48 (0.63–3.45)	0.365
Stage	I	1		1		1	
	II	3.5 (0.8–15)	0.095	4.28 (0.96–19.13)	0.057		
	III	5.0 (1.1–22)	0.003	5.67 (1.24–25.93)	0.025		
	IV	25.6 (6.0–109)	<0.001	18.58 (3.96–87.18)	<0.001		
Response to treatment	No	1		1		1	
	Yes	0.12 (0.07–0.21)	<0.001	0.16 (0.07–0.33)	<0.001		

TABLE 3 Viral loads obtained in the HPV16+ve group

	Level	N (%)	Viral Load median
All	Low VL (<12.3)	47 (32.41%)	3.85
	Medium VL (12.3–57)	50 (34.48%)	35.65
	High VL (>57)	48 (33.10%)	148.93
	All	145	60.80 (IQR 5.74–85)
Alive	Low VL (<12.3)	29 (27.88%)	4.11
	Medium VL (12.3–57)	33 (31.73%)	37.16
	High VL (>57)	42 (40.38%)	152.61
	All	104	74.57 (IQR 8.25–104.5)
Deceased	Low VL (<12.3)	18 (43.9%)	3.43
	Medium VL (12.3–57)	17 (41.46%)	32.72
	High VL (>57)	6 (14.63%)	123.25
	All	41	33.11 (IQR 3.6–43.5)

FIGURE 2 Kaplan–Meier survival curve stratified by viral load (Low, Medium and High). Survival time expressed in months from the diagnosis date. Data censored at 31st July 2020



#### 4 | DISCUSSION

To our knowledge this is the first study looking at viral load in anal cancer samples and its association with overall survival. Most anal cancers included in this study were positive for HPV (88.6%), with HPV16 being the clear dominant type (93.3%) in the positive cases. This is consistent with the high positivity of HPV in anal cancer and the high prevalence of HPV16 reported by Desanjosé et al.<sup>3</sup>

We have identified that HPV status (HPV positive) was associated with improved overall survival in the univariate analysis compared to HPV negative cases. When adjusted, HPV status continued to influence overall survival. This aligns with the systematic review by Urbute et al., where the authors found HPV DNA positive anal cancers

have significantly better OS compared with HPV negative cancers.<sup>55</sup> Moreover, this observation is consistent with an emerging pattern in other cancers associated with HPV, including cervical,<sup>43,44</sup> oropharyngeal,<sup>45,46</sup> penile<sup>47,48</sup> and vulvar cancers.<sup>49</sup>

The ddPCR assay indicated that high viral load as measured by quantifying HPV16 L1 gene copies was associated with a better clinical outcome than low copies of L1 in the univariate analysis, when compared with low and medium VL. However, when Cox HR was adjusted, viral load did not influence the overall survival. This could be due to the relatively small sample size; as the confidence interval just exceeds one, it is possible that a larger study may tip into significance. If we consider the unadjusted analysis, our observations with viral load and survival are



TABLE 4 Univariate and multivariate hazard ratio of L1 viral load derived using Cox regression (N = 145)

Variable	Level	Unadjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value	Adjusted HR (95% Cis)	p value
Viral Load	Low (<12.3)	1		1		1	
	Medium (12.3–57)	0.91 (0.47–1.76)	0.774	1.04 (0.45–2.40)	0.924	0.80 (0.40–1.60)	0.531
	High (>57)	0.28 (0.11–0.71)	0.007	0.39 (0.12–1.24)	0.111	0.27 (0.11–0.68)	0.006
Sex	Male	1		1		1	
	Female	1.2 (0.6–2.4)	0.625	1.09 (0.47–2.53)	0.838	1.35 (0.65–2.76)	0.419
Age	<50	1		1		1	
	50–59	1.51 (0.58–4.0)	0.398	0.77 (0.25–2.32)	0.639	1.30 (0.50–3.41)	0.588
	60–69	0.94 (0.34–2.6)	0.912	2.20 (0.69–7.01)	0.183	0.85 (0.30–2.40)	0.753
	70 and over	1.53 (0.56–4.2)	0.405	3.05 (0.757–12.31)	0.117	1.65 (0.58–4.65)	0.347
Stage	I	1		1			
	II	2.2 (0.48–10)	0.302	2.31 (0.48–11.18)	0.299		
	III	2.9 (0.62–14)	0.178	2.58 (0.50–13.23)	0.254		
	IV	25.2 (5.65–113)	<0.001	21.52 (4.01–115.41)	<0.001		
Response to treatment	No	1		1			
	Yes	0.13 (0.064–0.27)	<0.001	0.23 (0.09–0.56)	0.001		

similar to those seen for oropharynx, where high viral load correlated with improved survival using the same ddPCR technology applied in the present study.<sup>23</sup> Moreover, other investigators have shown a link with viral load and survival in cancers of the cervix,<sup>21,22</sup> head and neck<sup>23–27</sup> and anus.<sup>32</sup> Although use of viral load for active clinical management is still unclear; a biomarker of risk(s) that may inform multidisciplinary meeting discussion and/or serve as a marker for therapeutic strategies could have value.

The reason as to why higher viral load may plausibly confer a better prognosis (notwithstanding any association with stage) in HPV associated disease is not fully understood. It is feasible that a higher viral load in epithelial cells may maximise exposure to immune effector cells. Additionally, in cervical cancer cases, high viral load, mediated by integrated or episomal genomes has been shown to link with high levels of viral oncoprotein expression. For integrated genomes, epigenetic drivers and changes in stability of transcribed E6 E7 mRNAs are the main mechanisms of increased viral oncogene expression.<sup>50</sup> By contrast, episomal genomes may have the ability to express the entire viral proteome and this may repress viral oncogene expression. For oropharyngeal cancers, studies have shown that those cases that have an episomal genome status and high viral load have a more favourable prognosis.<sup>25,33–35</sup> We did not explore physical status of the genome in the present work, but this would be an interesting area to explore in the future.

We decided to use ddPCR given the relative lack of data on the implications of VL in the anal disease context and the fact that ddPCR delivers a high precision,<sup>23,51</sup>

arguably higher than that achievable by normalised real-time PCR. Additionally, ddPCR platforms are likely to have an increasing role in service laboratories to support precision testing in solid and liquid biopsies including for the measurement of circulating HPV and tumour DNA.<sup>52–54</sup>

There are limitations to the study—although the sample set was well annotated it was still relatively small. A larger study would have conferred greater power to investigate the impact of viral load, particularly in view of the various adjustments. Also, we did not account for differences between margin versus canal tumours or percentage of tumour captured as this information was not available to us. Additionally, due to the relationship of HIV with anal cancer development, knowledge of HIV status and markers of immune status/competence would have been a valuable addition to our data set.

Notwithstanding the above limitations, we would argue that assessment of HPV status in anal cancer cases and viral load is worthy of further investigation. High viral load of HPV in anal cancer serves as a proxy for improved survival and may have potential as biomarker—particularly at an early stage in clinical management where response to treatment is unknown. Larger series and studies which investigate this relationship further are welcome.

#### CONFLICT OF INTEREST

DG: Received gratis consumables from Seegene to support the HPV genotyping of the anal cancer samples of this study. KC: KC's institution has received research funding

or gratis consumables to support research from the following commercial entities in the last 3 years: Cepheid, Euroimmun, GeneFirst, SelfScreen, Hiantis, Seegene, Roche, Abbott and Hologic. All other authors have nothing to declare.

#### AUTHOR CONTRIBUTION

DG was involved in the lab testing, analysis of data, and drafted the manuscript. RG performed the data retrieval, and critical appraisal of the manuscript. AK and AS assisted with lab experiments and critical appraisal of the manuscript. JP assisted with the statistical analysis and critical appraisal of the manuscript. KK assisted with the statistical design. MTGH and SVG assisted in critical appraisal of the manuscript. KC was involved in planning, draft supervision and assisted in drafting and critical appraisal of the manuscript.

#### ETHICS APPROVAL STATEMENT

Favourable ethical opinion to conduct the research was provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482. This was further supported by approval for use of samples through the National Research for Scotland Bioresource (20/ES/0061), application reference SR1283.

#### PATIENT CONSENT STATEMENT

Not applicable.

#### DATA AVAILABILITY STATEMENT

Data in anonymized form can be made available upon reasonable request to the senior author, and following due process of governance and the Scottish Data Protection Regulations.

#### ORCID

Daniel Guerendain  <https://orcid.org/0000-0002-7536-1308>

Jiafeng Pan  <https://orcid.org/0000-0001-5993-3209>

#### REFERENCES

- Lin C, Franceschi S, Clifford GM. Human papillomavirus types from infection to cancer in the anus, according to sex and HIV status: a systematic review and meta-analysis. *Lancet Infect Dis*. 2018;18:198-206.
- Hartwig S, St Guily JL, Dominiak-Felden G, Alemany L, de Sanjosé S. Estimation of the overall burden of cancers, pre-cancerous lesions, and genital warts attributable to 9-valent HPV vaccine types in women and men in Europe. *Infect Agent Cancer*. 2017;12:19.
- de Sanjosé S, Serrano B, Tous S, et al. Burden of human papillomavirus (HPV)-related cancers attributable to HPVs 6/11/16/18/31/33/45/52 and 58. *JNCI Cancer Spectr*. 2019;2(4):pk045. doi:10.1093/jncics/pky045
- Islami F, Ferlay J, Lortet-Tieulent J, Bray F, Jemal A. International trends in anal cancer incidence rates. *Int J Epidemiol*. 2017;46(3):924-938. doi:10.1093/ije/dyw276
- Robinson D, Coupland V, Møller H. An analysis of temporal and generational trends in the incidence of anal and other HPV-related cancers in Southeast England. *Br J Cancer*. 2009;100(3):527-531. doi:10.1038/sj.bjc.6604871
- Cancer Research UK. Anal cancer incidence statistics. 2021. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/anal-cancer/incidence#heading=Two>
- Cancer Stat Facts: Anal Cancer. National cancer institute. [Online] Accessed September 01, 2021. <https://seer.cancer.gov/statfacts/html/anus.html>
- Colorectal Cancer (Includes Additional Cancer Types not on Website). Scotland: trends in incidence 1993-2017. Isdscotland.org. 2019. Isdscotland.org. Accessed July 24, 2019. Available from: <https://www.isdscotland.org>
- Tseng HF, Morgenstern H, Mack TM, Peters RK. Risk factors for anal cancer: results of a population-based case-control study. *Cancer Causes Control*. 2003;14(9):837-846.
- Valvo F, Ciurlia E, Avuzzi B, et al. Cancer of the anal region. *Crit Rev Oncol Hematol*. 2019;135:115-127. doi:10.1016/j.critrevonc.2018.12.007
- Lin C, Slama J, Gonzalez P, et al. Cervical determinants of anal HPV infection and high-grade anal lesions in women: a collaborative pooled analysis. *Lancet Infect Dis*. 2019;19(8):880-891. doi:10.1016/S1473-3099(19)30164-1
- Albuquerque A, Stockdale CK, Heller D, et al. Vulvar high-grade squamous intraepithelial lesions and cancer as a risk factor for anal cancer: a review. *J Low Genit Tract Dis*. 2022;26(1):32-37. doi:10.1097/LGT.0000000000000631. PMID: 34670242.
- Saleem AM, Paulus JK, Shapter AP, Baxter NN, Roberts PL, Ricciardi R. Risk of anal cancer in a cohort with human papillomavirus-related gynecologic neoplasm [published correction appears in *Obstet Gynecol*. 2013 Apr;121(4):881]. *Obstet Gynecol*. 2011;117(3):643-649. doi:10.1097/AOG.0b013e31820bf16
- Clifford GM, Georges D, Shiels MS, et al. A meta-analysis of anal cancer incidence by risk group: toward a unified anal cancer risk scale. *Int J Cancer*. 2021;148(1):38-47. doi:10.1002/ijc.33185
- Dandapani SV, Eaton M, Thomas CR Jr, Pagnini PG. HIV-positive anal cancer: an update for the clinician. *J Gastrointest Oncol*. 2010;1(1):34-44. doi:10.3978/j.issn.2078-6891.2010.005
- Hernández-Ramírez RU, Qin L, Lin H, et al. Association of immunosuppression and human immunodeficiency virus (HIV) viremia with anal cancer risk in persons living with HIV in the United States and Canada. *Clin Infect Dis*. 2020;70(6):1176-1185. doi:10.1093/cid/ciz329
- Glynn-Jones R, Nilsson PJ, Aschele C, et al. Anal Cancer: ESMO-ESSO-ESTRO clinical practice guidelines. *Ann Oncol*. 2014;25(suppl 3):iii10-iii20.
- Neibart SS, Manne SL, Jabbour SK. Quality of life after radiotherapy for rectal and anal cancer. *Curr Colorectal Cancer Rep*. 2020;16(1):1-10. doi:10.1007/s11888-019-00448-w
- Fakhrian K, Sauer T, Dinkel A, et al. Chronic adverse events and quality of life after radiochemotherapy in anal cancer patients. A single institution experience and review of the literature.

- Strahlenther Onkol.* 2013;189(6):486-494. doi:10.1007/s00066-013-0314-5
20. Das P, Cantor SB, Parker CL, et al. Long-term quality of life after radiotherapy for the treatment of anal cancer. *Cancer* vol. 2010;116(4):822-829. doi:10.1002/ncr.24906
  21. Kim JY, Park S, Nam BH, et al. Low initial human papilloma viral load implicates worse prognosis in patients with uterine cervical cancer treated with radiotherapy. *J Clin Oncol* 2009;27(30):5088-5093. doi:10.1200/JCO.2009.22.4659
  22. Deng T, Feng Y, Zheng J, Huang Q, Liu J. Low initial human papillomavirus viral load may indicate worse prognosis in patients with cervical carcinoma treated with surgery. *J Gynecol Oncol*. 2015;26(2):111-117. doi:10.3802/jgo.2015.26.2.111
  23. Stevenson A, Wakeham K, Pan J, et al. Droplet digital PCR quantification suggests that higher viral load correlates with improved survival in HPV-positive oropharyngeal tumours. *J Clin Virol*. 2020;129:104505. doi:10.1016/j.jcv.2020.104505
  24. Cohen MA, Basha SR, Reichenbach DK, Robertson E, Sewell DA. Increased viral load correlates with improved survival in HPV-16-associated tonsil carcinoma patients. *Acta Otolaryngol*. 2008;128(5):583-589. doi:10.1080/00016480701558880
  25. Mellin H, Dahlgren L, Munck-Wikland E, et al. Human papillomavirus type 16 is episomal and a high viral load may be correlated to better prognosis in tonsillar cancer. *Int J Cancer*. 2002;102(2):152-158. doi:10.1002/ijc.10669
  26. Hashida Y, Higuchi T, Matsumoto S, et al. Prognostic significance of HPV16 viral load level in patients with oropharyngeal cancer. *Cancer Sci*. 2021. doi:10.1111/cas.15105
  27. Biesaga B, Mucha-Malecka A, Janecka-Widla A, et al. Differences in the prognosis of HPV16-positive patients with squamous cell carcinoma of head and neck according to viral load and expression of P16. *J Cancer Res Clin Oncol*. 2018;144(1):63-73. doi:10.1007/s00432-017-25312
  28. Poizot-Martin I, Henry M, Benhaim S, Obry-Roguet V, Figarella D, Tamalet C. High level of HPV 16 and 18 DNA load in anal swabs from male and female HIV-1 infected patients. *J Clin Virol*. 2009;44(4):314-317. doi:10.1016/j.jcv.2009.02.003
  29. Pierangeli A, Scagnolari C, Degener AM, et al. Type-specific human papillomavirus-DNA load in anal infection in HIV-positive men. *AIDS*. 2008;22:1929-1935.
  30. Drobacheff C, Dupont P, Mougjin C, et al. Anal human papillomavirus DNA screening by hybrid capture IITM in human immunodeficiency virus-positive patients with or without anal intercourse. *Eur J Dermatol*. 2003;13:367-371.
  31. Mafusecka E, Chmielik E, Suwiński R, et al. Significance of HPV16 viral load testing in anal cancer. *Pathol Oncol Res*. 2020;26(4):2191-2199. doi:10.1007/s12253-020-00801-7
  32. Rödel F, Wieland U, Fraunholz I, et al. Human papillomavirus DNA load and p16INK4a expression predict for local control in patients with anal squamous cell carcinoma treated with chemoradiotherapy. *Int J Cancer*. 2015;136(2):278-288. doi:10.1002/ijc.28979
  33. Olthof NC, Straetmans JM, Snoeck R, Ramaekers FC, Kremer B, Speel EJ. Next-generation treatment strategies for human papillomavirus-related head and neck squamous cell carcinoma: where do we go? *Rev Med Virol*. 2012;22(2):88-105.
  34. Olthof NC, Huebbers CU, Kolligs J, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines.
  35. Rasmussen CL, Sand FL, Hoffmann Frederiksen M, Kaae Andersen K, Kjaer SK. Does HPV status influence survival after vulvar cancer? *Int J Cancer*. 2018;142(6):1158-1165. doi:10.1002/ijc.31139
  36. Moorghen M, Wong N. Standards and datasets for reporting cancers. Dataset for histopathological reporting of anal cancer. March 2018 The Royal College of Pathologists. 2018.
  37. National Records of Scotland. Mid-year population estimates: Scotland and its NHS Board areas, total population by sex: 1981 to 2019. <https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/population-estimates-time-series-data>
  38. Amin MB, Greene FL, Edge SB, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99. doi:10.3322/caac.21388
  39. Rao S, Guren MG, Khan K, et al. ESMO Guidelines Committee. ESMO Guidelines Committee. Anal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2021;32(9):1087-1100. doi:10.1016/j.annonc.2021.06.015
  40. Core Team R. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Austria; 2020. <https://www.R-project.org/>
  41. Therneau T. A package for survival analysis in R. R package version 3.2-10. 2021. URL: <https://CRAN.R-project.org/package=survival>
  42. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer; 2000. ISBN 0-387-98784-3.
  43. Li P, Tan Y, Zhu LX, et al. Prognostic value of HPV DNA status in cervical cancer before treatment: a systematic review and meta-analysis. *Oncotarget*. 2017;8(39):66352-66359. doi:10.18632/oncotarget.18558
  44. Nicolás I, Marimon L, Barnadas E, et al. HPV-negative tumors of the uterine cervix. *Mod Pathol*. 2019;32(8):1189-1196. doi:10.1038/s41379-019-0249-1
  45. Wakeham K, Pan J, Pollock KG, et al. A prospective cohort study of human papillomavirus-driven oropharyngeal cancers: implications for prognosis and immunisation. *Clin Oncol*. 2019. doi:10.1016/J.CLON.2019.05.010
  46. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010;363(1):24-35. doi:10.1056/NEJMoa0912217
  47. Sand FL, Rasmussen CL, Frederiksen MH, Andersen KK, Kjaer SK. Prognostic significance of HPV and p16 status in men diagnosed with penile cancer: a systematic review and meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2018;27(10):1123-1132. doi:10.1158/1055-9965.EPI-18-0322
  48. Chu C, Chen K, Tan X, et al. Prevalence of human papillomavirus and implication on survival in Chinese penile cancer. *Virchows Arch*. 2020;477(5):667-675. doi:10.1007/s00428-020-02831-7
  49. Lee LJ, Howitt B, Catalano P, et al. Prognostic importance of human papillomavirus (HPV) and p16 positivity in squamous cell carcinoma of the vulva treated with radiotherapy. *Gynecol Oncol*. 2016;142(2):293-298. doi:10.1016/j.ygyno.2016.05.019
  50. Jeon S, Lambert PF. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of

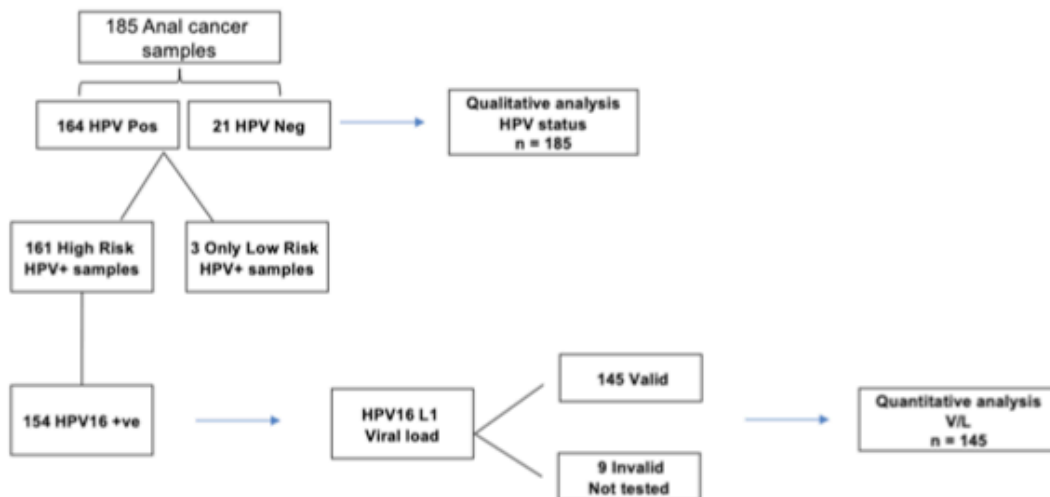


- E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc Natl Acad Sci USA*. 1995;92(5):1654-1658. doi:10.1073/pnas.92.5.1654
51. Zhang Y, Zhang Z, Wang Z, et al. Development of a droplet digital PCR assay for sensitive detection of porcine circovirus 3. *Mol Cell Probes*. 2019;43:50-57. doi:10.1200/JCO.19.02444. Epub 2020 Feb 4.
  52. Chera BS, Kumar S, Shen C, et al. Plasma circulating tumor HPV DNA for the surveillance of cancer recurrence in HPV-associated oropharyngeal cancer. *J Clin Oncol*. 2020;38(10):1050-1058. doi:10.1200/JCO.19.02444. Epub 2020 Feb 4. Erratum in: *J Clin Oncol*. 2020 Oct 20;38(30):3579. PMID: 32017652; PMCID: PMC7106982.
  53. Krasniqi E, Barba M, Venuti A, et al. Circulating HPV DNA in the management of oropharyngeal and cervical cancers: current knowledge and future perspectives. *J Clin Med*. 2021;10(7):1525. doi:10.3390/jcm10071525
  54. Jeannot E, Latouche A, Bonneau C, et al. Circulating HPV DNA as a marker for early detection of relapse in patients with cervical cancer. *Clin Cancer Res*. 2021;27(21):5869-5877. doi:10.1158/1078-0432.CCR-21-0625
  55. Urbute A, Rasmussen CL, Belmonte F, et al. Prognostic significance of HPV DNA and p16<sup>INK4a</sup> in Anal cancer: a systematic review and meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2020;29(4):703-710. doi:10.1158/1055-9965.EPI-19-1259

**How to cite this article:** Guerendiaín D, Grigorescu R, Kirk A, et al. HPV status and HPV16 viral load in anal cancer and its association with clinical outcome. *Cancer Med*. 2022;11:4193-4203. doi: [10.1002/cam4.4771](https://doi.org/10.1002/cam4.4771)

## APPENDIX A

Summary of study flow used, including the qualitative identification of HPV in anal cancer and the viral load investigation confined to HPV16 +ve only samples.





Article

# Mapping HPV 16 Sub-Lineages in Anal Cancer and Implications for Disease Outcomes

Daniel Guerendiain <sup>1,2,\*</sup>, Laila Sara Arroyo Mühr <sup>3</sup>, Raluca Grigorescu <sup>4</sup>, Matthew T. G. Holden <sup>2</sup> and Kate Cuschieri <sup>1</sup>

<sup>1</sup> Scottish HPV Reference Laboratory, Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh EH16 4SA, UK

<sup>2</sup> School of Medicine, University of St Andrews, St Andrews KY16 9TE, UK

<sup>3</sup> International HPV Reference Center, Department of Laboratory Medicine, Karolinska Institutet, 141 86 Stockholm, Sweden

<sup>4</sup> Department of Pathology, Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh EH16 4SA, UK

\* Correspondence: dgr7@st-andrews.ac.uk or daniel.guerendiain@nhslothian.scot.nhs.uk

**Abstract:** The incidence of anal cancer is rising worldwide. As identified in cervical cancer management, an improvement in the early detection and management of anal pre-cancer is essential. In other cancers associated with human papillomavirus (HPV), HPV 16 sub-lineages have been shown to be associated with disease status and prognosis. However, in anal cancer, they have been under-explored. A total of 119 HPV 16-positive anal cancer lesions diagnosed between 2009 and 2018 in Scotland and 134 HPV 16-positive residual rectal swabs from asymptomatic men collected in 2016/7 were whole genome sequenced. The association of HPV 16 sub-lineages with underlying disease status (cancer vs. asymptomatic) and overall survival in anal cancer samples was assessed (comparing A1 vs non-A1 sub-lineages). A1 was the dominant sub-lineage present in the anal cancer (76.5%) and the asymptomatic (76.1%) cohorts. A2 was the second most dominant sub-lineage in both groups (16.8% and 17.2%, respectively). We did not observe significant associations of sub-lineage with demographics, clinical variables or survival (A1 vs. non-A1 sub-lineages (HR 0.83, 0.28–2.46  $p = 0.743$ )). HPV 16 sub-lineages do not appear to cluster with disease vs asymptomatic carriage or be independently associated with outcomes in anal cancer patients. Further international studies on anal HPV sub-lineage mapping will help to determine whether this is a consistent observation.

**Keywords:** human papillomavirus; HPV 16 sub-lineages; anal cancer



**Citation:** Guerendiain, D.; Mühr, L.S.A.; Grigorescu, R.; Holden, M.T.G.; Cuschieri, K. Mapping HPV 16 Sub-Lineages in Anal Cancer and Implications for Disease Outcomes. *Diagnostics* **2022**, *12*, 3222. <https://doi.org/10.3390/diagnostics12123222>

Academic Editors: Fabio Bottari and Anna Daniela Iacobone

Received: 31 October 2022

Accepted: 9 December 2022

Published: 19 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Anal cancer is one of the six cancers shown to have a human papillomavirus (HPV) aetiology [1]. Most HPV-positive anal cancers are caused by HPV type 16 (HPV 16), and in a recent population-based assessment in Scotland, in cases diagnosed between 2009–2018, HPV 16 was detected in 93.3% of the HPV positive cases [2], higher than the amount of HPV 16 attributable to cervical cancer [3].

Additionally, as with other HPV-driven cancers, anal cancer incidence is increasing worldwide, including in the USA and Europe [4–7].

HPVs are formally classified as “types” based on the nucleotide sequence of the open reading frame (ORF) coding for the major capsid protein: L1 [8]. HPV types differ by more than 10% of their primary sequence compared to their most closely related type [8]. Phenotypic differences in HPV types with respect to disease risk and tissue tropism are well-established, and this knowledge has informed the development of effective vaccines and HPV-based cervical screening assays. However, below the level of HPV type exist lineages (with 2–10% variation) and sub-lineages (0.5% to 2% variation) [9], and the implications of this level of variation on clinical outcomes of infection is less established.

For HPV 16, four lineages have been identified (lineages A, B, C and D), as well as 16 sub-lineages: A, including A1–A3 (previously named European) and A4 (Asian) sub-lineages; B, including B1 (African-1, Afr1a) and B2 (African-1, Afr1b), B3 and B4 sub-lineages; C1 (African-2, Afr2a), C2, C3 and C4; and D, including D1 (North American, NA1), D2 (Asian-American, AA2), D3 (Asian-American, AA1) and D4 sub-lineages [9].

Although some investigators have assessed the global distribution of sub-lineages, the majority have focused on cervical cancers rather than other HPV-driven cancers. In 2013, Cornet et al. looked at the HPV lineages in cervical cancers and showed that European sub-lineages (A1–A3) were the most common in all regions of the world, except in sub-Saharan Africa and East Asia, whereas the African sub-lineages dominated in the northern sub-Saharan region of Africa, and the Asian variant in East Asia [10]. Nicolás-Párraga et al. (2016) found similar results, with A1–3 present in 95.65% of the cases in Europe, 78.26% in Central/South America (D in 21.73%) and 80% in Asia (12% A4 and 7.69% D) [11].

In terms of HPV 16 sub-lineages present in the anus, data is relatively sparse. Volpini et al. (2017) investigated the HPV 16 variants in anal samples collated in Brazil, finding that 70.8% were classified as A1–3 sub-lineages and 29.2% as “other” [12]. A recent systematic review, performed by Ferreira et al. (2021) of genetic variants of HPV-16 in men, found HPV 16 lineages vary according to anatomical and geographical regions, but they found that European samples had a high prevalence (86.59%) of HPV 16 lineage A [13].

In the context of cervical disease, evidence suggests that sub-lineages and variants may be independently associated with poor clinical outcomes. Mirabello et al. (2015) [14] revealed a higher risk of disease associated with B/C/D lineages as a group compared to the A lineage. Clifford et al. (2019). also found an increased cervical cancer risk for A3, A4 and D-(sub-) lineages vs the A1 sub-lineage. A more recent study by Lang Kuhs et al. (2022) looked into the genetic variation of HPV 16 and its association with clinical outcomes in HPV 16-positive oropharyngeal cancer patients. They investigated different high-risk single nucleotide polymorphisms (SNPs) and found that those with one or more high-risk SNPs had a median survival time of 3.96 years compared to 18.67 years for those with no high-risk SNPs. Most of these SNPs were common to the D2 sub-lineage, which have also been associated with higher risk of cancer in the cervix [14]. However equivalent studies on anal cancer are rare.

We recently identified that the viral load of HPV16 in anal cancer may be informative for prognosis [2]. Now, due to the information published on the association of HPV 16 sub-lineages and cancer risk yet the comparative absence of data in the anal context, we aimed to better understand the pattern and dominance of HPV 16 sub-lineages in a population-based cohort of anal cancer and to determine whether significant associations with sub-lineage and demographic or clinical variables existed. Data obtained from the cancer cohort was contextualized and compared to variant profile in anal samples obtained from an asymptomatic population.

## 2. Material and Methods

### 2.1. Sample Collection

A total of 150 HPV 16-positive anal cancers and 182 DNA extracts from residual rectal swabs obtained from asymptomatic men were selected for HPV 16 sub-lineage identification through whole genome sequencing (WGS).

#### 2.1.1. Anal Cancer Cohort, Collection and Annotation

For the present work, we used the same anal cancer ( $n = 150$ ) sample set as described in detail in Guerendiain et al. (2022) [2]. Briefly, nucleic acid extract associated with archived formalin-fixed, paraffin-embedded tissue was genotyped using the Seegene Anyplex II 28 (Seoul, Korea), followed by storage at  $-80\text{ }^{\circ}\text{C}$ . Anal cancer biopsy samples were taken between 2009 and 2018 as part of the management of patients with anal disease from 3 of the 14 territorial health boards in Scotland (NHS Lothian, NHS Borders and NHS Fife).

HPV typing was performed at the Scottish HPV Reference Laboratory, Edinburgh, UK. One 10 µm section per sample was obtained and incubated in Seegene Universal Lysis Buffer (LB) at 65 °C overnight. DNA extraction was performed using the Microlab Nimbus IVD (Hamilton, Reno, USA) with the StarMAg Universal cartridge Kit (Seegene), following manufacturers' instructions. Mastermix was prepared with the Nimbus and PCR on the CFX Real-Time PCR instrument (Biorad, CA, USA).

As described in Guerendiain et al. (2022), clinico-demographic information was obtained in January 2020, specifically the patient's age, sex, stage of cancer (using the American Joint Committee on Cancer (AJCC) TNM system) [15], response to treatment, date of diagnosis and vital (dead/alive) status. Age and stage of cancer were considered at the time of diagnosis. Vital status information and date of death data was censored in July 2020.

Cases categorized according to the various clinical and demographic variables are summarized in Table 1. Age was stratified in 4 different groups: <50, 50–59, 60–69 and ≥0. Response to treatment was organized in 3 groups: yes, no or unknown, following the ESMO guidelines for anal cancer [16]. Cancer stage was aggregated in 5 groups: I, II, III, IV and unknown, following the AJCC system effective January 2018 [15].

**Table 1.** Anal cancer cohort: clinical characteristics and demographics with valid NGS analysis.

Variable	Level	n = 119	%
Sex	Female	90	75.6
	Male	29	24.4
Age	<50	15	12.6
	50–59	32	26.9
	60–69	39	32.8
	70 and over	33	27.7
Stage	I	18	15.1
	II	48	40.3
	III	36	30.2
	IV	16	13.4
	Unknown	1	0.8
Response to treatment	Yes	95	79.8
	No	17	14.3
	Unknown	7	5.9
Vital status	Alive	87	73.1
	Deceased	30	25.2
	Unknown	2	1.7

### 2.1.2. Residual Rectal Swabs from Asymptomatic Men

To contextualize the sequences observed in the anal cancers, a disease-free control group of anonymized residual rectal swabs obtained from asymptomatic men attending sexual health clinics were collated for downstream WGS. These samples had previously been genotyped as a consequence of immunization surveillance in Scotland [17]. DNA was extracted from the residual rectal swabs by Qiagen MDx (Hilden, Germany) or Seegene Universal Extraction System, obtaining an eluate volume of 100 µL. HPV genotyping was performed using the Seegene Anyplex II and the Optiplex HPV Genotyping Kit (Heidelberg, Germany), detecting 28 and 24 different HPV types, respectively.

A total of 182 anonymized DNA extracts from residual rectal swabs were collated. Samples had originally been collected in the years 2016–2017 and were included if positive for HPV 16.

## 2.2. Governance

Use of samples for the present project was approved by the Southeast of Scotland National Research for Scotland Bioresource (NRS) (application reference SR 1283 and SR1364). A favorable ethical opinion to conduct the research was provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482.

## 2.3. PCR Target-Enrichment for Deep Sequencing of HPV 16

HPV 16 whole genome material was amplified using 47 overlapping amplicons described in Cullen et al. [18] and optimized by Arroyo et al. [19]. Briefly, primer sets were divided into five different reactions to decrease self-dimer and cross-primer dimer formation. PCRs were performed using Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany) and 0.2  $\mu$ M of each primer, according to manufacturers' instructions. PCR amplification products were pooled together according to sample name prior to library preparation.

## 2.4. Library Preparation

Libraries were prepared using the Illumina DNA prep kit (San Diego, CA, USA) following the manufacturer's instructions, using 450 ng of DNA in 35  $\mu$ L as input. Sequencing was performed using the Illumina MiSeq instrument and the Illumina MiSeq reagent kit v2 500 cycles (2  $\times$  250 bp). Libraries were normalized to 4 nM in combination with 12.5 pM of PhiX (Illumina).

## 2.5. Quality Control and Quality Analysis

HPV 16-positive (SiHa) and HPV-negative (water) controls were added at the DNA extraction step and carried through the PCR, library preparation and data analysis stages. Individual amplification products were assessed using a Bioanalyzer (Santa Clara, CA, USA). Quality control for library preparation included both controlling the library size using an Agilent TapeStation (Santa Clara, CA, USA) and determining the DNA concentration using the Qubit dsDNA High-Sensitivity Assay Kit (ThermoFisher, Waltham, MA, USA).

As a further quality control for analysis of the sequence data generated, a subset of 25 fastq files were sent to the International HPV Reference Laboratory in Karolinska, Sweden for independent bioinformatic analysis and sub-lineage identification.

## 2.6. Bioinformatic Analysis

Reads obtained from Illumina were de-multiplexed and converted to fastq files. All fastq files were quality and adaptor trimmed using Trimmomatic (v0.39) [20]. Only high-quality paired reads (-phred 33 -leading 3 -trailing 3 -slidingWindow: 4:15) with 150 bp were used for further analysis. FASTQC tools were further used to assess whether any adaptors remained [21]. High-quality reads were then mapped to the HPV 16 reference genome from the Papillomavirus Episteme (PaVE) [22] using bwa (v0.7.17) [23], to create a sam file. Due to the circular HPV genome, the reference genome was modified by adding the 258 nucleotides from the beginning to the end of the genome sequence to not lose coverage of amplicons 46 and 47. SAMtools (v1.14) [24] was then used to convert files from sam to bam and to curate files for the variant calling. BCFtools (v1.14), mpileup and consensus tools were used for the variant calling and for the generation of a consensus sequence [25], using default parameters. Positions not covered were annotated as Ns.

New consensus files were aligned using MAFFT (v7.490) with default parameters [26]. A manual edit was performed when required. Maximum likelihood trees were inferred using RaxML (v2.0.8) [27] with the GTR substitution model (ML + transfer bootstrap expectation + consensus, 1 run, 100 reps). Visualization of the trees generated by RaxML was performed using Figtree (v1.4.4). Each sample was assigned with a sub-lineage corresponding to the nearest neighbor.

Sub-lineages references were obtained from the PAVE for each of the HPV 16 sub-lineages: A1 (K02718.1), A2 (AF536179.1), A3 (HQ644236.1), A4 (AF534061.1), B1 (AF536180.1), B2 (HQ644298.1), B3 (HQ644298.1), B4 (KU053914.1), C1 (AF472509.1), C2 (HQ644244.1),



C3 (KU053920.1), C4 (KU053925.1), D1 (HQ644257.1), D2 (AY686579.1), D3 (AF402678.1) and D4 (AF402678.1) A sub-lineage assignment was performed for all specimens excluding those with  $<100\times$  median depth or low genome coverage ( $<80\%$  genome coverage).

### 2.7. Assessment of Variants According to Clinic-Demographic Characteristics and Survival Analysis

To assess the relationship between HPV sub-lineages and different factors (two or more independent variables), a univariate logistic regression analysis was performed between HPV 16 sub-lineages (HPV 16 A1-positive vs. HPV 16 non-A1-positive), age at diagnosis, collection year and health board of diagnosis. Adjustment was performed for age group ( $<50$ , 50–59, 60–69 and 70 or over), sex, response to treatment, stage of cancer and vital status (dead or alive). Comparison was performed between A1 vs. non-A1 sub-lineages due to the small number of samples identified from the different sub-lineages. The non-A1 group includes the following sub-lineages: A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3 and D4.

Odds ratios (OR) were calculated to quantify the strength of the association between HPV 16 sub-lineages and the demographic and clinical data. All the statistics were obtained using R-studio macOS, (version 1.2.1335) [28]. The distribution of sub-lineages in anal cancers vs. the asymptomatic population was assessed with sequences from the two groups displayed in a phylogenetic tree.

Overall survival by HPV 16 sub-lineages (HPV 16 A1-positive vs. HPV 16 non-A1-positive) was analyzed using the Kaplan-Meier method. The univariate and multivariate hazard ratios of HPV 16 sub-lineages (HPV 16 A1-positive vs. HPV 16 non-A1-positive) for all-cause death were derived using the cox proportional hazard model. A univariate and multivariate model was derived; age ( $<50$ , 50–59, 60–69, 70+), sex, stage (I, II, III, IV) and response to treatment (no, yes) were adjusted for. All the statistical analyses were performed using R-studio (version 1.2.1335) [29]. Differences in prevalence of the HPV 16 sub-lineages between the anal cancer cohort and asymptomatic control cohort are presented descriptively.

## 3. Results

A total of 182 asymptomatic/control samples and 150 anal cancer samples were subjected to WGS. In the anal cancer cohort, 119/150 (79.3%) samples passed the quality parameters, and in the asymptomatic men cohort, 134/182 (73.6%) were valid. This left a total of 253 samples for inclusion for detailed sequencing/phylogenetic analysis. Twenty-five of these sequences were also analyzed by the International HPV Reference Center as a further quality control for analysis. Results showed 100% agreement.

### 3.1. Distribution of HPV 16 Sub-Lineages in Anal Cancers

Of the 119 cancer cases with sufficient read depth ( $>100\times$  median depth and  $>80\%$  genome), the HPV 16 sub-lineage A1 was identified in 91 anal cancer samples (76.5%), followed by A2, which was identified in 20/119 (16.8%) of samples. A4 was detected in 5/119 samples (4.2%). Two samples were classified as B1 (1.7%), one as A3 (0.8%) and one as D1 (0.8%). Further detail of HPV 16 sub-lineages in anal cancers is described in detail in Table 2 and Figure 1.

**Table 2.** HPV 16 sub-lineages identified in the anal cancer and asymptomatic cohorts.

Sub-Lineage	Anal Cancer		Asymptomatic Group	
	N	% (N = 119)	N	% (N = 134)
HPV 16 A1	91	76.5%	102	76.1%
HPV 16 A2	20	16.8%	23	17.2%
HPV 16 A3	1	0.8%	0	0%
HPV 16 A4	5	4.2%	0	0%

Table 2. Cont.

Sub-Lineage	Anal Cancer		Asymptomatic Group	
	N	% (N = 119)	N	% (N = 134)
HPV 16 B1	2	1.7%	2	1.5%
HPV 16 B2	0	0.0%	1	0.7%
HPV 16 B3	0	0.0%	0	0%
HPV 16 B4	0	0.0%	0	0%
HPV 16 C1	0	0.0%	2	1.5%
HPV 16 C2	0	0.0%	0	0%
HPV 16 C3	0	0.0%	0	0%
HPV 16 C4	0	0.0%	0	0%
HPV 16 D1	1	0.8%	4	3.0%
HPV 16 D2	0	0.0%	0	0%
HPV 16 D3	0	0.0%	0	0%
HPV 16 D4	0	0.0%	0	0%
<b>Total</b>	<b>119</b>		<b>134</b>	

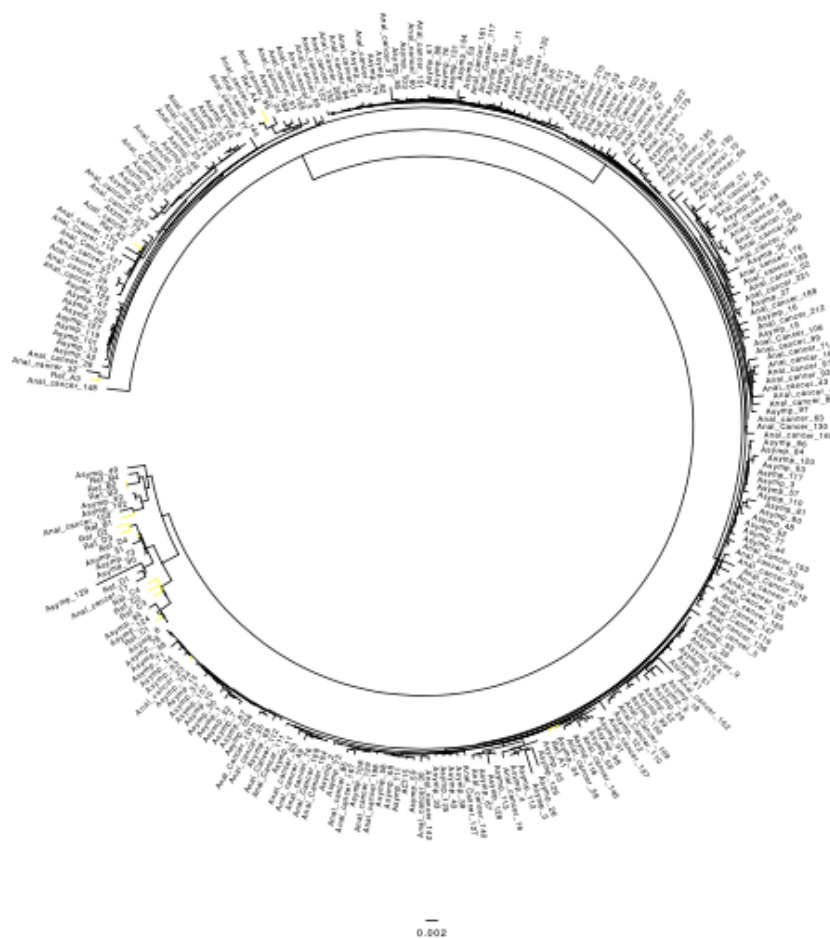


Figure 1. Phylogenetic tree representing the HPV 16 sub-lineages present in the anal sample and control groups.

### 3.2. HPV 16 Sub-Lineages in the Control Cohort

Of the 134 control samples, most samples were classified (76.1%) as A1, followed by A2, identified in 23/134 (17.2%) of samples. D1 sub-lineage was identified in 4/134 samples (3.0%), and C1 and B1 were identified in two cases each (1.5%). B2 was present in 1/134 (0.7%). Table 2 describes the number of cases identified for each sub-lineage, and Figure 1 contains the phylogenetic tree obtained from the control cohort.

### 3.3. Differences in Prevalence of HPV 16 Sub-Lineages between Anal Cancer and Control Cohort

No major differences in the proportion of A1 and A2 sub-lineages were observed between the case vs. control cohorts, being 76.4% vs. 76.0% and 16.2% vs. 17.0%, respectively. Whereas the case cohort revealed the presence of the A4 sub-lineage in 4.2% of anal cancers, this sub-lineage was not present in the control cohort. Conversely, the C lineage was present in 1.5% of control specimens and was not detected in cancer cases. Finally, there was a slight increase in the D1 sub-lineage observed in the control vs. case cohort (3.0% vs. 0.8%).

### 3.4. Association of HPV 16 Sub-Lineages with Demographic and Clinical Variables

From the 119 anal cancers, four samples did not contain vital status information and were not included in the analysis. Due to the dominance of the A1 sub-lineage, the logistic analysis and odds ratio analysis were performed based on the presence or absence of the HPV 16 sub-lineage A1.

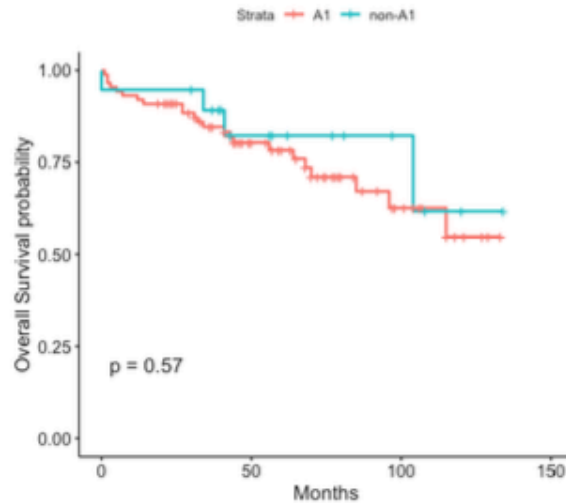
No significant differences in A1 positivity with sex, age, response to treatment, stage or vital status were observed. This observation was consistent for the adjusted analysis (Table 3).

**Table 3.** Influence of A1 sub-lineage presence stratified by demographic and clinical variables. The comparator for the odds ratio (univariate and adjusted) is A1 absence.

Variable	Level	Unadjusted OR (95% CIs)	p Value	Adjusted OR (95% CIs)	p Value
Sex	Male	1		1	
	Female	1.12 (0.39–2.92)	0.827	1.09 (0.37–3.00)	0.87
Age	<50	1		1	
	50–59	1.20 (0.27–4.80)	0.8	1.11 (0.24–4.56)	0.89
	60–69	1.50 (0.34–5.93)	0.57	1.82 (0.40–7.67)	0.416
	70 and over	1.37 (0.30–5.67)	0.666	1.63 (0.34–7.41)	0.529
Response to treatment	No	1		1	
	Yes	1.02 (0.26–3.26)	0.968	1.18 (0.34–7.41)	0.528
Stage	I	1		1	
	II	1.36 (0.37–4.62)	0.625	1.28 (0.33–4.51)	0.706
	III	1.60 (0.40–6.11)	0.486	1.56 (0.38–6.06)	0.522
	IV	1.80 (0.36–10.40)	0.478	3.03 (0.42–29.47)	0.289
Vital Status	Alive	1		1	
	Deceased	1.01 (0.39–2.85)	0.983	0.92 (0.25–3.81)	0.907

### 3.5. HPV 16 Sub-Lineages and Overall Survival

For the Kaplan-Meier estimator, overall survival was calculated by classifying HPV 16 sub-lineages into A1 presence or absence. No differences in overall survival were found between both sub-lineage groups ( $p = 0.57$ ), Figure 2.



**Figure 2.** Kaplan-Meier survival curve stratified by HPV 16 sub-lineages (A1 vs. Non-A1). Survival time expressed in months from the diagnosis date. Data censored on 31 July 2020.

Table 4 shows overall survival stratified by the clinical and demographic variables (age group, sex, cancer stage and response to treatment), with HPV 16 sub-lineages categorized into the two groups (A1 vs. non-A1) with A1 as the reference. Non-A1 (vs. A1) was not associated with improved overall survival in the univariate analysis, with a hazard ratio (HR) of 0.87 (0.37–2,  $p = 0.751$ ). Variables associated with worse overall survival in the univariate model were stage IV vs. stage I with HR of 15.7 (3.38–72.8),  $p < 0.001$  and response to treatment vs. no response to treatment with HR of 0.11 (0.05–0.25)  $p < 0.001$ . After adjustment for age, gender, stage and response to treatment, non-A1 sub-lineages did not significantly influence the overall survival compared to A1, with a HR 0.83 (0.28–2.46,  $p = 0.743$ ).

**Table 4.** Hazard ratio of HPV 16 sub-lineages (univariate and multivariate) derived from Cox regression ( $N = 115$ ) in anal cancer samples collected between 2009 to 2018 in the southeast of Scotland.

Variable	Level	Unadjusted HR (95% Cis)	p Value	Adjusted HR (95% Cis)	p Value
HPV 16 sub-lineage	A1 ( $n = 88$ )	1		1	
	Non-A1 ( $n = 27$ )	0.87 (0.37–2)	0.751	0.83 (0.28–2.46)	0.743
Sex	Male	1		1	
	Female	1.2 (0.48–2.9)	0.71	0.88 (0.32–2.39)	0.795
Age	<50	1		1	
	50–59	1.10 (0.33–3.70)	0.877	0.83 (0.21–3.26)	0.788
	60–69	0.85 (0.26–2.8)	0.795	2.67 (0.607–11.72)	0.194
	70 and over	1.54 (0.48–5.0)	0.466	5.56 (1.082–28.58)	0.04
Stage	I	1		1	
	II	1.7 (0.37–8.1)	0.49	2.34 (0.47–11.74)	0.302
	III	2.4 (0.50–11.6)	0.274	2.26 (0.42–12.27)	0.344
	IV	15.7 (3.38–72.8)	<0.001	15.95 (2.45–103.82)	0.004
Response to treatment	No	1		1	
	Yes	0.11 (0.05–0.25)	<0.001	0.12 (0.03–0.39)	<0.001



If only A1 and A2 samples were considered, no significant differences in HR were found when using A1 as the reference (HR 0.74, 0.25–2.1,  $p = 0.575$ ).

### 3.6. Integration

Although it was not the main aim of the study, the absence of part of the HPV 16 genome was identified in 13/119 (10.92%) of the anal cancer samples. This absence indicates the potential integration of the HPV16 in the human genome. The E2 gene was the most frequently missing region, followed by E4, E5 and L2 E1 and L1 (see Table 5 for details). Notably, all cases retained E6 and E7 oncogenes. Due to the small number of cases in which integration was detected, no further analysis was performed. No integration was detected in the asymptomatic cohort.

**Table 5.** HPV integration identified in the anal cancer cohort.

HPV Genes Integration in the Anal Cancer Samples ( $n = 13$ )	N
L1 only	1
E1 only	1
E1, E2, E4	1
E2, E5, part E2	1
E2, E4, E5, L2, L1	3
E1, E2, E4, E5 and part L2	2
E1, E2, E4, E5, L2 and part L1	3
E1, E2, E4, E5, L2 and L1 complete	1

## 4. Discussion

Previously, we described that 93.3% of HPV-positive anal cancer cases diagnosed in Scotland between 2019 and 2018 were caused by HPV 16. In this study, we have identified that 76% of cases belonged to the A1 sub-lineage, followed by A2 (16%).

In the control group of asymptomatic men, a similar prevalence of A1 and A2 was observed. Differences identified were the presence of A4 in the anal cancers (4.7%), which was absent in the control group; presence of the C lineage only detected in the control group; and the presence of the sub-lineage D1 in the control group (3%), which had a lower prevalence of 0.81% in the cancers. This higher prevalence of sub-lineages A1 and A2 is consistent with previously published studies in European cohorts. Gonçalves et al. (2022) found a higher prevalence of the A lineage in the anal canal of asymptomatic men, mainly A1 [29], and Nicolás-Párraga et al. (2016) found that A1–3 sub-lineages were identified in 96.1% of the European cases [30]. Beyond Europe, Volpini et al. (2017) investigated the HPV 16 variants in cervical and anal samples collated in Brazil and determined that proportionally less of the anal cancer samples (70.8%) were classified as A1–3 sub-lineages [12].

The data collated in the study add to the limited information on the pattern and implications of HPV sub-lineages in the anus. Though we did not see significant associations with demographic and underlying disease status, these observations need to be confirmed or refuted by future studies with larger sample sizes.

To our knowledge, no other studies have investigated the association of HPV 16 sub-lineages in anal cancer and overall survival. We did not observe that A1 vs. non-A1 sub-lineages influenced overall survival in the univariate and adjusted analysis. Interestingly, a recent study was reported by Lang Kuhs et al. (2022) in which the authors looked into the genetic variation of HPV 16 and its association with clinical outcomes in HPV 16-positive oropharyngeal cancer patients [31]. They investigated different high-risk single nucleotide polymorphisms (SNPs) and found that those with one or more high-risk SNPs had significantly shorter median survival times. Most of these SNPs were common to the D2 sub-lineage, which has also been associated with a higher risk of cancer in the cervix [14]. Due to the absence of D2 cases in the present study, we were not able to explore

this in the present work; however, the identification of these high-risk SNPs may be very helpful for patient and treatment management.

Although we identified potential integration of the HPV16 genome (calculated through the loss of the sequence), due to the small number, we did not perform any further analysis, including in relation to implications for survival. Given the relative lack of information on the extent and implications of integration in anal cancer, we would assert that this is an area that would benefit from further study.

We acknowledge this study has limitations; the asymptomatic population were all men, whereas the cancer population had a majority of female (75.63%) samples compared to males (24.37%); this was due to pragmatic reasons relating to available material. However, data did not show differences in the distribution of HPV 16 sub-lineages between women and men in the anal cancer group. Additionally, as discussed earlier, we believe the observations made in the present work would benefit from validation in a larger sample of cases and controls and would hope this study serves as a primer for such. Though the number of cases of cancers was not trivial ( $n = 253$ ), particularly given that the Scottish European age-standardized rate (EASR) (per 100,000 person-years at risk) was 2.6 in 2017, we appreciate that detecting rarer sub-lineages with precision can take large sample sizes.

In the UK, there is no screening program for anal cancer. However, since 2017, there has been an opportunistic vaccination program for MSM, and in 2019, the national HPV vaccination became gender neutral. In term of vaccines, a study from Godi et al. (2019) reported that HPV 16 lineage variants B, C and D exhibited slightly (<two-fold) reduced sensitivity to nonavalent vaccine sera compared to lineage A [32].

Therefore, the high prevalence of lineage A in the samples included in this study could be interpreted as positive for vaccine efficacy, particularly given that gender-neutral vaccination is now a part of core policy in the UK and several other countries.

This study has demonstrated the technical feasibility of detecting HPV 16 sub-lineages in anal cancer samples and residual material from rectal swabs. Though some differences in the presence of non-A sub-lineages were detectable between the cancer and asymptomatic population, the consistency, magnitude and implications of these would benefit from further study. The domination of lineage A is consistent with existing European data and suggests that sub-lineage identification in itself may not be informative for prognostication.

**Author Contributions:** D.G. was involved in the planning of experiments, delivered the end-to-end whole genome sequencing process and performed data analysis, including the analysis of next-generation sequencing data. D.G. also drafted the manuscript. L.S.A.M. assisted with the planning of laboratory experiments, supporting with quality checking/analysis of data, including sequencing data, and performing critical appraisal of the manuscript. R.G. performed original data retrieval on the clinical cohort, including the collation of clinical-demographic variables, and supported critical appraisal of the manuscript. M.T.G.H. was the lead academic supervisor for the project and was involved in advising on experimental methodology and technology, providing support for data analysis and performing critical appraisal of the manuscript. K.C. was the principal clinical investigator for the project and supported with interaction with the bio-resource and pathology team for sample collation, advising on experimental and analytical methodology, providing support for data analysis and assisting in the drafting and critical appraisal of the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Use of samples for the present project was approved by the Southeast of Scotland National Research for Scotland Bioresource (NRS) (application reference SR 1283 (24 September 2019) and SR1364 (22 January 2020)). Favorable ethical opinion to conduct the research was provided by University of St Andrews Teaching and Research Ethics Committee, reference MD 14482 (5 July 2019).

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Request for data in anonymized form can be made available upon reasonable request to the senior author and following due process of governance and the Scottish Data Protection Regulations. GenBank submission IDs 2637056 and 2638666.

**Conflicts of Interest:** D.G.: Received gratis consumables from Seegene to support the HPV genotyping of the anal cancer samples. K.C.: K.C.'s institution has received research funding or gratis consumables to support research from the following commercial entities in the last 3 years: Cepheid, Euroimmun, GeneFirst, SelfScreen, Hiantis, Seegene, Roche, Abbott and Hologic. All other authors have nothing to declare.

### Abbreviations

Human papillomavirus (HPV), hazard ratio (HR), European age-standardized rate (EASR), human immunodeficiency virus (HIV), deoxyribonucleic acid (DNA), overall survival (OS), National Health Service (NHS), whole genome sequencing (WGS).

### References

- De Sanjosé, S.; Serrano, B.; Tous, S.; Alejo, M.; Lloveras, B.; Quirós, B.; Clavero, O.; Vidal, A.; Ferrándiz-Pulido, C.; Pavón, M.; et al. Burden of Human Papillomavirus (HPV)-Related Cancers Attributable to HPVs 6/11/16/18/31/33/45/52 and 58. *JNCI Cancer Spectr.* **2018**, *2*, pky045. [CrossRef] [PubMed]
- Guerendiain, D.; Grigorescu, R.; Kirk, A.; Stevenson, A.; Holden, M.T.G.; Pan, J.; Kavanagh, K.; Graham, S.V.; Cuschieri, K. HPV status and HPV16 viral load in anal cancer and its association with clinical outcome. *Cancer Med.* **2022**, *11*, 4193–4203. [CrossRef] [PubMed]
- Cuschieri, K.; Brewster, D.; Williams, A.R.W.; Millan, D.; Murray, G.; Nicoll, S.; Imrie, J.; Hardie, A.; Graham, C.; Cubie, H.A. Distribution of HPV types associated with cervical cancers in Scotland and implications for the impact of HPV vaccines. *Br. J. Cancer* **2010**, *102*, 930–932. [CrossRef]
- Islami, F.; Ferlay, J.; Lortet-Tieulent, J.; Bray, F.; Jemal, A. International trends in anal cancer incidence rates. *Int. J. Epidemiol.* **2017**, *46*, 924–938. [CrossRef] [PubMed]
- Robinson, D.; Coupland, V.; Moller, H. An analysis of temporal and generational trends in the incidence of anal and other HPV-related cancers in Southeast England. *Br. J. Cancer* **2009**, *100*, 527–531. [CrossRef] [PubMed]
- Anal Cancer Incidence Statistics | Cancer Research, UK. Available online: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/anal-cancer/incidence#heading-Two> (accessed on 11 September 2021).
- Anal Cancer—Cancer Stat Facts. Available online: <https://seer.cancer.gov/statfacts/html/anus.html> (accessed on 11 September 2021).
- De Villiers, E.-M.; Fauquet, C.; Broker, T.R.; Bernard, H.-U.; zur Hausen, H. Classification of papillomaviruses. *Virology* **2004**, *324*, 17–27. [CrossRef]
- Burk, R.D.; Harari, A.; Chen, Z. Human papillomavirus genome variants. *Virology* **2013**, *445*, 232–243. [CrossRef]
- Cornet, I.; Gheit, T.; Iannacone, M.R.; Vignat, J.; Sylla, B.S.; Del Mistro, A.; Franceschi, S.; Tommasino, M.; Clifford, G.M.; on behalf of the IARC HPV Variant Study Group. HPV16 genetic variation and the development of cervical cancer worldwide. *Br. J. Cancer* **2012**, *108*, 240–244. [CrossRef]
- Clifford, G.M.; Tenet, V.; Georges, D.; Alemany, L.; Pavón, M.A.; Chen, Z.; Yeager, M.; Cullen, M.; Boland, J.F.; Bass, S.; et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res.* **2019**, *7*, 67–74. [CrossRef]
- Volpini, L.P.B.; Boldrini, N.A.T.; de Freitas, L.B.; Miranda, A.E.; Spano, L.C. The high prevalence of HPV and HPV16 European variants in cervical and anal samples of HIV-seropositive women with normal Pap test results. *PLoS ONE* **2017**, *12*, e0176422. [CrossRef]
- Ferreira, M.T.; Gonçalves, M.G.; López, R.V.M.; Sichero, L. Genetic variants of HPV-16 and their geographical and anatomical distribution in men: A systematic review with meta-analysis. *Virology* **2021**, *558*, 134–144. [CrossRef] [PubMed]
- Mirabello, L.; Yeager, M.; Cullen, M.; Boland, J.F.; Chen, Z.; Wentzensen, N.; Zhang, X.; Yu, K.; Yang, Q.; Mitchell, J.; et al. HPV16 Sublineage Associations with Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **2016**, *108*, djw100. [CrossRef] [PubMed]
- American Joint Committee on Cancer. *AJCC Cancer Staging Manual*, 8th ed.; Springer: Berlin/Heidelberg, Germany, 2017; p. 275. Available online: [www.cancerstaging.org/ajcc@facs.org](http://www.cancerstaging.org/ajcc@facs.org) (accessed on 11 September 2021).
- Glynn-Jones, R.; Nilsson, P.; Aschele, C.; Goh, V.; Peiffert, D.; Cervantes, A.; Arnold, D. Anal cancer: ESMO-ESSO-ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2014**, *25*, iii10–iii20. [CrossRef] [PubMed]
- Cameron, R.L.; Cuschieri, K.; Pollock, K.G.J. Baseline HPV prevalence in rectal swabs from men attending a sexual health clinic in Scotland: Assessing the potential impact of a selective HPV vaccination programme for men who have sex with men. *Sex. Transm. Infect.* **2019**, *96*, 55–57. [CrossRef] [PubMed]

18. Cullen, M.; Boland, J.F.; Schiffman, M.; Zhang, X.; Wentzensen, N.; Yang, Q.; Chen, Z.; Yu, K.; Mitchell, J.; Roberson, D.; et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* **2015**, *1*, 3–11. [[CrossRef](#)] [[PubMed](#)]
19. Arroyo-Mühr, L.S.; Lagheden, C.; Hultin, E.; Eklund, C.; Adami, H.-O.; Dillner, J.; Sundström, K. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: Prospective population-based study. *Br. J. Cancer* **2018**, *119*, 1163–1168. [[CrossRef](#)] [[PubMed](#)]
20. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
21. Babraham Bioinformatics—FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 18 April 2022).
22. PaVE. Available online: [https://pave.niaid.nih.gov/explore/reference\\_genomes/human\\_genomes](https://pave.niaid.nih.gov/explore/reference_genomes/human_genomes) (accessed on 31 October 2022).
23. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows—Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)]
24. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
25. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993. [[CrossRef](#)]
26. Kazutaka, K.; Misakwa, K.; Kei-ichi, K.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
27. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [[CrossRef](#)] [[PubMed](#)]
28. Team, R. Rstudio: Integrated Development for r. Rstudio, pbc, Boston, MA. 2020. Available online: <http://www.Rstudio.Com> (accessed on 3 March 2022).
29. Gonçalves, M.G.; Ferreira, M.T.; López, R.V.M.; Ferreira, S.; Sirak, B.; Baggio, M.L.; Lazcano-Ponce, E.; Nyitray, A.G.; Giuliano, A.R.; Villa, L.L.; et al. Prevalence and persistence of HPV-16 molecular variants in the anal canal of men: The HIM study. *J. Clin. Virol.* **2022**, *149*, 105128. [[CrossRef](#)] [[PubMed](#)]
30. Nicolás-Párraga, S.; Gandini, C.; Pimenoff, V.N.; Alemany, L.; Sanjosé, S.; Bosch, F.X.; Bravo, I.G.; the RIS HPV TT and HPV VVAP Study Groups. HPV16 variants distribution in invasive cancers of the cervix, vulva, vagina, penis, and anus. *Cancer Med.* **2016**, *5*, 2909–2919. [[CrossRef](#)] [[PubMed](#)]
31. Kuhs, K.L.; Faden, D.; Chen, L.; Smith, D.; Pinheiro, M.; Wood, C.; Davis, S.; Yeager, M.; Boland, J.; Cullen, M.; et al. Genetic variation within the human papillomavirus type 16 genome is associated with oropharyngeal cancer prognosis. *Ann. Oncol.* **2022**, *33*, 638–648. [[CrossRef](#)] [[PubMed](#)]
32. Godi, A.; Kemp, T.J.; Pinto, L.A.; Beddows, S. Sensitivity of Human Papillomavirus (HPV) Lineage and Sublineage Variant Pseudoviruses to Neutralization by Nonavalent Vaccine Antibodies. *J. Infect. Dis.* **2019**, *220*, 1940–1945. [[CrossRef](#)] [[PubMed](#)]



## 12. Ethical Approval Documents

### Lothian NRS Bioresource Approvals

Document Name	QF-TGU-A-SAMREQA	VERSION 1.1	Page	1 of 2	Review date	25-Sep-2019
---------------	------------------	-------------	------	--------	-------------	-------------

#### LOTHIAN NRS BIORESOURCE SAMPLE REQUEST ANSWER FORM

Sample Request number:	SR1243
Name of Researcher:	Daniel Guerendiain
Address of Researcher:	Scottish HPV Reference Lab Department of Laboratory Medicine Royal Infirmary of Edinburgh EH16 4SA
Study Title:	PhD Project: Taxonomic and functional diversity of HPV variants in an era of vaccination.
Ethical status:	15/ES/0094
Material Requested	<p>Release and use of the following is approved by Tissue Governance for the purposes of the above study. Approval is given on the basis of the information submitted on the request form and accompanying emails.</p> <p>200 cervical liquid-based cytology samples annotated with HPV status, underlying pathology and vaccination status which are currently stored in the Scottish HPV Archive.</p> <p>Sections of 100 oropharyngeal cancer cancers sent to the Scottish HPV Reference laboratory for routine HPV testing. Underlying pathology information and survival outcome as a consequence of a clinical audit.</p> <p>100 CIN3+ (FFPE blocks or derivates as available)</p> <p>Linked clinical; data to include underlying pathology grade, age (not date of birth) and immunisation status. Survival outcome on the oropharyngeal cases (as a consequence of a clinical audit) .</p> <p>The above data to be accessed and gathered by named members of the clinical team who are not involved in the research.</p>

#### REQUEST AUTHORISED

Date:	29-May-2019
Authorised by:	

Author	: Frances Rae	Date	: 25-Sep-2015
Authority for Issue	: Craig Marshall	Date	: 25-Sep-2015
Quality Checked	: Craig Marshall	Date	: 25-Sep-2015

Document Name	QF-TGU-A-SAMREQA	VERSION 1.1	Page	2 of 2	Review date	25-Sep-2019
---------------	------------------	-------------	------	--------	-------------	-------------

**REQUEST REJECTED**

Date:	
Authorised by:	
Reason	

Author	: Frances Rae	Date	: 25-Sep-2015
Authority for Issue	: Craig Marshall	Date	: 25-Sep-2015
Quality Checked	: Craig Marshall	Date	: 25-Sep-2015

Document Name	QF-TGU-A-SAMREQA	VERSION 1.1	Page	1 of 2	Review date	25-Sep-2021
---------------	------------------	-------------	------	--------	-------------	-------------

## LOTHIAN NRS BIORESOURCE SAMPLE REQUEST ANSWER FORM

Sample Request number:	SR1283 (Amendment 07-Sep-2020)
Name of Researcher:	Daniel Guerendiain
Address of Researcher:	Scottish HPV Reference Laboratory Royal Infirmary of Edinburgh 51 Little France Crescent Edinburgh EH16 4SA
Study Title:	Taxonomic and functional diversity of HPV type and variants in high grade anal lesions and cancers.
Ethical status:	20/ES/0061 Previously 15/ES/0094
Material Requested	<p>Release and use of the following archival tissue and associated pseudonymised data is approved by Tissue Governance for the purposes of the above study. Approval is given on the basis of the information submitted on the request form.</p> <p>Material from 100 anal intraepithelial neoplasia (AIN2&amp; 3 samples) and 100 anal cancers diagnosed between 2009-2018</p> <p>Data to be provided and pseudonymised by Drs Paul Fineron and Raluca Grigorescu.</p> <p>Data release is restricted to:</p> <ul style="list-style-type: none"> <li>• Pathology grade</li> <li>• Age (not date of birth)</li> <li>• Date of diagnosis</li> <li>• Morphology</li> <li>• Treatment</li> <li>• Recurrence and time to recurrence</li> <li>• Smoking status</li> <li>• Vital status (Dead/Alive)</li> </ul> <p><b>Approval amended (07-Sep-2020)</b> to send nucleic acid extracts from approximately 188 HPV positive anal samples from this study for ddPCR testing at the University of Glasgow (Professor Sheila Graham's group) to be tested by a very sensitive assay they have developed to identify the viral load of HPV in oropharyngeal samples (droplet digital PCR (ddPCR)).</p>

Author	: Frances Rae	Date	: 25-Sep-2015
Authority for Issue	: Craig Marshall	Date	: 25-Sep-2015
Quality Checked	: Craig Marshall	Date	: 25-Sep-2015

<b>Document Name</b>	QF-TGU-A-SAMREQA	VERSION 1.1	Page	2 of 2	Review date	25-Sep-2021
----------------------	------------------	-------------	------	--------	-------------	-------------

	Approval will run to the 1 <sup>st</sup> of October 2022. If this timescale needs to be extended for any reason, Tissue Governance must be notified otherwise a new application will be required.
--	---

**REQUEST AUTHORISED**

Date:	07-Sep-2020
Authorised by:	

**REQUEST REJECTED**

Date:	
Authorised by:	
Reason	

<b>Author</b>	: Frances Rae	<b>Date</b>	: 25-Sep-2015
<b>Authority for Issue</b>	: Craig Marshall	<b>Date</b>	: 25-Sep-2015
<b>Quality Checked</b>	: Craig Marshall	<b>Date</b>	: 25-Sep-2015



Document Name	QF-TGU-A-SAMREQA	VERSION 1.1	Page	1 of 2	Review date	25-Sep-2021
---------------	------------------	-------------	------	--------	-------------	-------------

### LOTHIAN NRS BIORESOURCE SAMPLE REQUEST ANSWER FORM

Sample Request number:	SR1364
Name of Researcher:	Daniel Guerendiain
Address of Researcher:	Scottish HPV Reference Laboratory Royal Infirmary of Edinburgh 51 Little France Crescent Edinburgh EH16 4SA
Study Title:	Taxonomic and functional diversity of HPV types and variants in rectal samples from asymptomatic males
Ethical status:	15/ES/0094
Material Requested	<p>Release and use of the following samples and associated pseudonymised data is approved by Tissue Governance for the purposes of the above study. Approval is given on the basis of the information submitted on the request form.</p> <p>A maximum of 300 residual rectal swabs collected as part of the HPV male surveillance program and previously used in SR716.</p> <p>Data as described on the request form to be linked via the study ID only without further accessing patient records.</p> <p>Approval will run to the 1<sup>st</sup> of October 2022. If this timescale needs to be extended for any reason, Tissue Governance must be notified otherwise a new application will be required.</p>

#### REQUEST AUTHORISED

Date:	22-Jan-2020
Authorised by:	

#### REQUEST REJECTED

Date:	
Authorised by:	

Author	: Frances Rae	Date	: 25-Sep-2015
Authority for Issue	: Craig Marshall	Date	: 25-Sep-2015
Quality Checked	: Craig Marshall	Date	: 25-Sep-2015

Document Name	QF-TGU-A-SAMREQA	VERSION 1.1	Page	2 of 2	Review date	25-Sep-2021
---------------	------------------	-------------	------	--------	-------------	-------------

Reason	
--------	--

Author	: Frances Rae	Date	: 25-Sep-2015
Authority for Issue	: Craig Marshall	Date	: 25-Sep-2015
Quality Checked	: Craig Marshall	Date	: 25-Sep-2015

# St Andrews Ethical Approvals



School of Medicine Ethics Committee

14 October 2019

Daniel Guerendiain  
School of Medicine

Dear Mr Guerendiain

Thank you for submitting your ethical amendment application.

The School of Medicine Ethics Committee has approved this ethical amendment application:

<b>Original Approval Code:</b>	MD14482	<b>Original Approval Date:</b>	5 July 2019
<b>Amendment 1 Approval Date:</b>	14 October 2019	<b>Approval Expiry Date:</b>	5 July 2024
<b>Project Title:</b>	Taxonomic and functional diversity of HPV variants in an era of vaccination		
<b>Researcher(s):</b>	Daniel Guerendiain	<b>Supervisor/PI:</b>	Prof. Matthew Holden
<b>School/Unit:</b>	School of Medicine		

The following supporting documents are also acknowledged and approved:

1. Ethical Amendment Form
2. External Approval from NHS NRS Bioresource

This approval does not extend the originally granted approval period. If you require an extension to the approval period, you can write to your School Ethics Committee who may grant a discretionary extension of no greater than 6 months. For longer extensions, or for any further changes, you must submit an additional ethical amendment application. For all extensions, you should inform the School Ethics Committee when your study is complete.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
  - the details provided in your ethical amendment application (and the original ethical application where still relevant)
  - the University's [Principles of Good Research Conduct](#)
  - the conditions of any funding associated with your work
- that you obtain all applicable additional documents and approvals (see [the relevant webpage](#) for guidance) before research commences.

School of Medicine Ethics Committee

Dr Morven Shearer, SEC Convenor/Gill Rhodes, SEC Administrator  
School of Medicine, University of St Andrews, North Haugh, St Andrews, Fife. KY16 9TF  
T: 01334 461733 E: medethic@st-andrews.ac.uk  
The University of St Andrews is a charity registered in Scotland: No SC013532



University of  
St Andrews | FOUNDED |  
1413 |

School of Medicine Ethics Committee

You should retain this approval letter with your study paperwork.

Yours sincerely

Dr Morven Shearer  
Convenor of the School of Medicine Ethics Committee

cc. Prof. Matthew Holden

School of Medicine Ethics Committee

Dr Morven Shearer, SEC Convenor/Gill Rhodes, SEC Administrator  
School of Medicine, University of St Andrews, North Haugh, St Andrews, Fife. KY16 9TF  
T: 01334 461733 E: [medethic@st-andrews.ac.uk](mailto:medethic@st-andrews.ac.uk)  
The University of St Andrews is a charity registered in Scotland: No SC013532

School of Medicine Ethics Committee

06 February 2020

Daniel Guerendiain Regalado  
School of Medicine

Dear Daniel

Thank you for submitting your ethical amendment application.

The School of Medicine Ethics Committee has approved this ethical amendment application:

<b>Original Approval Code:</b>	MD14482	<b>Original Approval Date:</b>	5 July 2019
<b>Amendment 2 Approval Date:</b>	30 January 2020	<b>Approval Expiry Date:</b>	5 July 2024
<b>Project Title:</b>	Taxonomic and functional diversity of HPV variants in an era of vaccination		
<b>Researcher(s):</b>	Daniel Guerendiain Regalado	<b>Supervisor/PI:</b>	Professor Matthew Holden
<b>School/Unit:</b>	School of Medicine		

The following supporting documents are acknowledged and approved:

1. Ethical Amendment Form
2. Lothian NRS sample request form

This approval does not extend the originally granted approval period. If you require an extension to the approval period, you can write to your School Ethics Committee who may grant a discretionary extension of no greater than 6 months. For longer extensions, or for any further changes, you must submit an additional ethical amendment application. For all extensions, you should inform the School Ethics Committee when your study is complete.

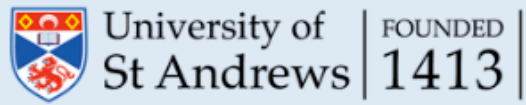
You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
  - the details provided in your ethical amendment application (and the original ethical application where still relevant)
  - the University's [Principles of Good Research Conduct](#)
  - the conditions of any funding associated with your work
- that you obtain all applicable additional documents and approvals (see [the relevant webpage](#) for guidance) before research commences.

School of Medicine Ethics Committee

Dr Morven Shearer, SEC Convenor/Gill Rhodes, SEC Administrator  
School of Medicine, University of St Andrews, North Haugh, St Andrews, Fife. KY16 9TF  
T: 01334 461733 E: medethic@st-andrews.ac.uk  
The University of St Andrews is a charity registered in Scotland: No SC013532



School of Medicine Ethics Committee

You should retain this approval letter with your study paperwork.

Yours sincerely

Dr Morven Shearer  
Convenor of the School of Medicine Ethics Committee

cc. Professor Matthew Holden

School of Medicine Ethics Committee  
Dr Morven Shearer, SEC Convenor/Gill Rhodes, SEC Administrator  
School of Medicine, University of St Andrews, North Haugh, St Andrews, Fife. KY16 9TF  
T: 01334 461733 E: medethic@st-andrews.ac.uk  
The University of St Andrews is a charity registered in Scotland: No SC013532

